Transportation Systems Engineering:
Theory and Methods

Applied Optimization

Volume 49

Series Editors:

Panos M. Pardalos
University of Florida, U.S.A.

Donald Hearn
University of Florida, U.S.A.

The titles published in this series are listed at the end of this volume.

Transportation Systems Engineering: Theory and Methods

by

Ennio Cascetta

Università degli Studi di Napoli "Federico II",
Dipartimento di Ingegneria dei Trasporti "Luigi Tocchetti",
Napoli, Italy

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4757-6875-6 ISBN 978-1-4757-6873-2 (eBook)
DOI 10.1007/978-1-4757-6873-2

Printed on acid-free paper

to Manuela, Vittorio and Priscilla

CONTENTS

About the Author XIII

Introduction XV

1. TRANSPORTATION SYSTEMS 1

1.1. Definition 1
1.2. Transportation system identification 5
1.2.1. Relevant spatial and supply characteristics 6
1.2.2. Relevant components of transport demand 9
1.2.3. Relevant temporal dimensions 12
1.3. Modeling transportation systems 19
Reference Notes 22

2. TRANSPORTATION SUPPLY MODELS 23

2.1. Introduction 23
2.2. Congested network models 24
2.2.1. Graph models 25
2.2.2. Performance variables and transportation costs 26
2.2.3. Flows 31
2.2.4. Link performance and cost functions 33
2.2.5. Impacts and impact functions 35
2.2.6. General formulation 36
2.3. Applications of Transportation Supply models 37
2.3.1. Supply models for continuous service systems 39
2.3.1.1. Graph models 39
2.3.1.2. Performance and cost functions 43
2.3.2. Supply models for scheduled service systems 58
2.3.2.1. Line-based graph models 59
2.3.2.2. Performance and cost functions 61
2.A. Review of Traffic Flow Theory 65
2.A.1. Models for running links 65
2.A.1.1. Fundamental variables 65
2.A.1.2. Stationary models 68
2.A.1.3. Non-stationary models 75
2.A.2. Models for queuing links 78
2.A.2.1. Fundamental variables 78
2.A.2.2. Deterministic models 79
2.A.2.3. Stochastic models 85
2.A.3. Application to signalized intersections 87
Reference Notes 92

3. RANDOM UTILITY THEORY 95

3.1. Introduction 95
3.2. Basic assumptions 96

3.3. Some random utility models 101
3.3.1. The Multinomial Logit model 101
3.3.2. The Single-Level Hierarchical Logit model 106
3.3.3. The Multi-Level Hierarchical Logit model* 113
3.3.4. The Cross-Nested Logit model* 122
3.3.5. The Generalized Extreme Value (GEV) model* 126
3.3.6. The Probit model 128
3.3.7. The Hybrid Logit-Probit model* 136
3.4. Choice set modeling* 137
3.5. Expected Maximum Perceived Utility and mathematical properties of random utility models 141
3.6. Direct and cross elasticities of random utility model 147
3.7. Aggregation methods for random utility models. 151
3.A. Derivation of logit models from the GEV model 157
3.A.1. Derivation of the Multinomial Logit model 157
3.A.2. Derivation of the Single-Level Hierarchical Logit model 158
3.A.3. Derivation of the Multi-Level Hierarchical Logit model 160
3.A.4. Derivation of the Cross-Nested Logit model 163
3.B. Random variables relevant for random utility models 165
3.B.1. The Gumbel random variable 165
3.B.2. The Multivariate Normal random variable 168
Reference Notes 169
4. TRANSPORTATION DEMAND MODELS 175
4.1. Introduction 175
4.2. Trip demand model systems 178
4.3. Examples of trip demand models 184
4.3.1. Emission or trip frequency models 184
4.3.2. Distribution models 188
4.3.3. Mode choice models 192
4.3.4. Path choice models 197
4.3.4.1. Path choice models for road systems 197
4.3.4.2. Path choice models for transit systems 207
4.3.5. A system of demand models 215
4.4. Trip-chaining demand models* 220
4.5. Applications of demand models 228
4.6. Freight transport demand models* 230
4.6.1. Multiregional Input-Output (MRIO) models 232
4.6.2. Freight mode choice models 243
Reference Notes 245
5. MODELS FOR TRAFFIC ASSIGNMENT TO TRANSPORTATION NETWORKS 251
5.1. Introduction 251
5.2. Definitions, assumptions, and basic equations 255
5.2.1. Supply model 256
5.2.2. Demand model 259
5.2.3. Feasible path and link flow sets 264
5.2.4. Network performance indicators 265
5.3. Models for assignment to Uncongested Networks 268

5.3.1. Models for Stochastic Uncongested Network assignment 270
5.3.2. Models for Deterministic Uncongested Network assignment 271
5.4. Rigid demand Users Equilibrium assignment models 274
5.4.1. Stochastic User Equilibrium models 276
5.4.2. Deterministic User Equilibrium models 281
5.4.3. Relationship between stochastic and deterministic equilibrium flows 288
5.4.4. System optimal assignment models 291
5.5. Assignment models with pre-trip/en-route path choice 297
5.6. Elastic demand User Equilibrium assignment models* 307
5.6.1. Single-mode assignment models 311
5.6.1.1. Elastic demand single-mode Stochastic User Equilibrium models 313
5.6.1.2. Elastic demand single-mode Deterministic User Equilibrium models 316
5.6.2. Multi-mode assignment models 321
5.7. Multi-class assignment models* 324
5.7.1. Differentiated congestion multi-class assignment models 327
5.7.2. Undifferentiated congestion multi-class assignment models 328
5.8. Inter-period Dynamic Process assignment models* 331
5.8.1. Definitions, assumptions and basic equations 332
5.8.1.1. Supply model 332
5.8.1.2. Demand model 333
5.8.1.3. Approaches to Dynamic Process modeling 336
5.8.2. Deterministic Process models 339
5.8.3. Stochastic Process models 345
5.9. Synthesis and application issues of assignment models 348
5.A. Optimization models for stochastic assignment 357
5.A.1. Stochastic Uncongested Network assignment 357
5.A.2. Stochastic User Equilibrium assignment 357
Reference Notes 360
6. INTRA-PERIOD (WITHIN-DAY) DYNAMIC MODELS* 367
6.1. Introduction 367
6.2. Supply models for continuous service systems 368
6.2.1. Continuous flow supply models 370
6.2.1.1. Variables and consistency conditions 371
6.2.1.2. Link performance and travel time functions 379
6.2.1.3. Path performance and travel time functions 380
6.2.1.4. Dynamic Network Loading models 384
6.2.1.5. Formalization of the overall supply model 387
6.2.2. Discrete flow supply models 388
6.2.2.1. Variables and consistency conditions 389
6.2.2.2. Link performance and travel time functions 393
6.2.2.3. Path performance and travel time functions 394
6.2.2.4. Dynamic Network Loading models 395
6.2.2.5. Formalization of the overall supply model 397
6.3. Demand models for continuous service systems 398
6.4. Demand-supply interaction models for continuous service systems 403
6.4.1. Uncongested network assignment models 403
6.4.2. User Equilibrium assignment models 406
6.4.3. Dynamic Process assignment models 410
6.5. Models for scheduled service systems 416

6.5.1. Models for regular low-frequency services 418
6.5.1.1. Supply models 419
6.5.1.2. Demand models 423
6.5.1.3. Demand-supply interaction models 425
6.5.2. Models for irregular high-frequency services 425
6.5.2.1. Supply models 425
6.5.2.2. Demand models 427
6.5.2.3. Demand-supply interaction models 432
Reference Notes 433
7. ALGORITHMS FOR TRAFFIC ASSIGNMENT TO TRANSPORTATION NETWORKS 435
7.1. Introduction 435
7.2. Shortest path algorithms 436
7.3. Algorithms for Uncongested Network assignment 440
7.3.1. Stochastic Uncongested Network assignment without explicit paths enumeration 441
7.3.1.1. SUN assignment with Logit path choice model 441
7.3.1.2. SUN assignment with Probit path choice model 448
7.3.2. Deterministic Uncongested Network assignment without explicit paths enumeration 453
7.4. Algorithms for rigid demand User Equilibrium assignment 456
7.4.1. Rigid demand Stochastic User Equilibrium 457
7.4.2. Rigid demand Deterministic User Equilibrium 461
7.4.3. Algorithms for System Optimal Assignment 466
7.5. Algorithms for assignment with pre-trip/en-route path choice 467
7.5.1. Shortest hyperpath algorithms 467
7.5.2. Algorithms for Uncongested Network assignment with pre-trip/en-route path choice 472
7.5.3. Algorithms for rigid demand User Equilibrium assignment with pre-trip/en-route path choice 473
7.6. Extensions of User Equilibrium assignment algorithms* 475
7.7. Applicative issues of assignment algorithms 481
Reference Notes 482
8. ESTIMATION OF TRAVEL DEMAND FLOWS 485
8.1. Introduction 485
8.2. Direct estimation of present demand 486
8.2.1. Sampling surveys 486
8.2.2. Sampling estimators 488
8.3. Disaggregate estimation of demand models 492
8.3.1. Model specification 493
8.3.2. Model calibration 494
8.3.3. Model validation 502
8.4. Disaggregate estimation of demand models with Stated Preferences surveys* 508
8.4.1. Definitions and types of survey 509
8.4.2. Survey design 511
8.4.3. Model calibration 517
8.5. Estimation of O-D demand flows using traffic counts 522

8.5.1. Maximum Likelihood and GLS estimators* ... 528
8.5.2. Bayesian estimators* ... 533
8.5.3. Applicative issues ... 535
8.5.4. Solution methods ... 538
8.6. Aggregate calibration of demand models using traffic counts ... 542
8.7. Estimation of intra-period dynamic demand flow using traffic counts* ... 548
8.7.1. Simultaneous Estimators ... 553
8.7.2. Sequential Estimators ... 554
8.8. Applications of demand estimation methods ... 555
8.8.1. Estimation of present demand ... 555
8.8.2. Estimation of demand variations (forecasting) ... 557
Reference Notes ... 559

9. TRANSPORTATION SUPPLY DESIGN MODELS ... 565

9.1. Introduction ... 565
9.2. General formulations of the Supply Design Problem ... 569
9.3. Some applications of Supply Design models ... 572
9.3.1. Models for road network layout design ... 572
9.3.2. Models for road network capacity design ... 576
9.3.3. Models for transit network design ... 577
9.3.4. Models for pricing design ... 581
9.3.5. Models for mixed design ... 583
9.4. Some algorithms for Supply Design Models ... 584
9.4.1. Algorithms for the discrete SDP ... 584
9.4.2. Algorithms for the continuous SDP ... 591
Reference Notes ... 596

10. TRANSPORTATION SYSTEMS ENGINEERING FOR PLANNING AND EVALUATION ... 599

10.1. Introduction ... 599
10.2. Transportation systems engineering and the decision-making process ... 600
10.3. Some areas of application ... 604
10.4. Evaluation of transportation system projects ... 606
10.4.1. Identification of relevant impacts ... 607
10.4.2. Identification and estimation of impact indicators ... 609
10.4.3. Computation of impacts perceived by the users ... 611
10.5. Methods for the comparison of alternative projects ... 625
10.5.1. Benefit-Cost analysis ... 626
10.5.2. Multi-Criteria analysis ... 632
Reference Notes ... 640

A. REVIEW OF NUMERICAL ANALYSIS ... 643

A.1. Sets and functions ... 643
A.1.1. Elements of set topology ... 643
A.1.2. Differentiable functions ... 645
A.1.3. Convex functions ... 649
A.2. Solution algorithms ... 652
A.3. Fixed point problems ... 652
A.3.1. Properties of fixed points ... 654
A.3.2. Solution algorithms for fixed point problems ... 656

A.4. Optimization problems 658
A.4.1. Properties of minimum points 659
A.4.1.1. Properties of minimum points on open sets 659
A.4.1.2. Properties of minimum points on closed sets 659
A.4.2. Solution algorithms for optimization problems 660
A.4.2.1. Mono-dimensional optimization algorithms 660
A.4.2.2. Unconstrained multi-dimensional optimization algorithms 664
A.4.2.3. Bounded variables multi-dimensional optimization algorithms 669
A.4.2.4. Linearly constrained multi-dimensional optimization algorithms 670
A.5. Variational inequality problems 673
A.5.1. Properties of variational inequalities 674
A.5.2. Solution algorithms for variational inequality problems 675
References **677**
Index **693**
Main Variables **705**

ABOUT THE AUTHOR

Ennio Cascetta, born in Naples in 1953, is a full professor of Transportation System Theory at the University of Naples "Federico II".

He has been responsible for several research projects at the national and international level including, recently, directorship of the Second Special Project on Transportation by the Italian National Research Council (CNR).

His experience as a teacher and researcher includes appointments at several universities, such as MIT and the University of Montreal. In addition to his academic and research work, dr. Cascetta currently acts as Associate Editor of the international journal, Transportation Science, and holds positions in many international scientific and professional organizations such as transportation Research C, WCTR, TRB and ATBR.

Dr. Cascetta's areas of expertise are analysis, modeling and estimation of transportation demand, static and dynamic assignment models, planning and pricing of transportation networks, traffic theory, road safety, ITS and control of urban traffic systems. He has published several books in Italian and over one hundred papers on his research in national and international journals and proceedings.

Dr. Cascetta activities outside academia also include a current appointment as transportation secretary of Campania Region and extensive consulting work related to the planning and management of national, regional and urban transportation systems.

INTRODUCTION

Science is made of facts just as a house
is made of bricks, but a collection of facts is no
more science than a pile of bricks is a house.
Henry Poincarè

The aim of the disciplines of praxis
is not theoretical knowledge .... It is to change
the forms of action ... .
Aristotle

Transportation systems engineering is a broad discipline aimed at the functional design of physical and/or organizational projects relating to transportation supply systems. These projects define the functional characteristics and performances of system elements (services, prices, infrastructures, vehicles, control, etc.) that, taken as a whole, provide transportation opportunities to satisfy the travel demand of persons and goods in a given area. The basic approach of transportation system engineering is to define the main characteristics of transportation services starting with the analysis and simulation of the demand for such services. Physical elements of the system are designed and/or identified among those available to provide the characteristics and performances required by the transportation services. Transportation system projects have to be technically feasible and defined on the basis of the quantitative evaluation of their main effects with respect to the objectives and constraints of the project itself. In the context of this general definition, there are projects of very different kinds. Such projects include the functional design of new infrastructures; the assessment of long-term investment programs; the evaluation of project financing schemes; the definition of schedules and pricing policies for transportation services; the definition of circulation and regulation schemes for urban road networks; the design of strategies for new advanced traffic control and information systems. In the proposed perspective the term "engineering" should be intended rather broadly. The models and techniques described in this book are often used by transportation system analysts with backgrounds in several disciplines such as urban planning, transportation economics, spatial system analysis and control engineering.

The difficulty, but also the fascination, of this professional practice derives from the intrinsic complexity of transportation systems. These are, in fact, "internally" complex systems, made up of many elements influencing each other both directly and indirectly, often non-linearly, with many feedback cycles. Furthermore, only some elements in the system are "technical" (vehicles, infrastructures, etc.), governed by the laws of physics and, as such, traditionally studied by engineers. On the other hand, the mechanisms underlying the functionality and the performances of these elements are often connected to travel demand and users' behavior. Thus the analysis of travel demand plays a central role in understanding and designing

transportation systems; however, it requires a different approach making reference to concepts traditionally used in social and economic sciences.

Apart from the internal complexity, transportation systems are closely interrelated with other systems which, from the point of view adopted, can be defined as external. Transport projects may have implications for the economy, the location and intensity of the activities in a given area, the environment, the quality of life and social cohesion. In short, they have a bearing on many, often contrasting, interests, as can easily be seen from the heated arguments that accompany almost all decisions concerning transport. The intensity of these impacts as well as the sensitivity to them, have grown considerably in recent decades as a result of the economic and social development of our civilization and have to be addressed in the design and evaluation of transportation projects.

For all these reasons, the consequences of projects cannot be predicted on the basis of pure experience and intuition. The latter, although prerequisites for any good design, do not allow a quantitative evaluation of the effects of a project and may be misleading for complex systems. In fact, simulations sometimes provide unexpected and apparently paradoxical results: a new infrastructure which increases congestion on existing facilities; local projects whose effects propagate to remote parts of the network; price increases which lead to revenues reductions; measures meant to reduce car usage which result in an overall increase in air pollution and energy consumption; and so on. Furthermore, the large number of design variables and the complexity of their interactions often require models and algorithms capable of simulating the effects of several combinations of such variables to help the designer to find satisfactory combinations. Finally, social fairness can only be addressed through a quantitative approach.

In order to develop solutions to these problems, the mathematical theory of transportation systems presented in this book has been developed over recent decades. This discipline is systematic in its approach. It is concerned with the relationships among the elements making up a transportation system and their performances. It is based on an autonomous theoretical nucleus and on analysis and calculation techniques derived from the contributions of many other disciplinary areas, especially economics, econometrics, and operation research, in addition to those traditionally more directly relevant for transportation engineers, such as traffic engineering, transportation infrastructures engineering and vehicles mechanics.

The discipline's theoretical foundation is, in my opinion, a "topological-behavioral" paradigm consisting of a set of assumptions and a limited number of functional relationships. This paradigm represents in an abstract way transportation services and their performances (supply model), travel demand and behavior of system users (demand model) as well as their interactions (demand/supply interaction model).

Over the years, these assumptions and relationships have been extended and formalized. The general mathematical properties of these models have been investigated producing a wide and internally consistent system of results; these results possess a certain degree of formal elegance and can be applied to models

differing in their mathematical formulation and their basic assumptions. This does not exclude new and significant theoretical and methodological developments. In fact, this is probably one of the areas of system engineering, in which research is most active, able to generate extensions and generalizations internally, and able to widen and even replace the assumptions on which it is based. Examples can be seen in research developments on interactions of transportation with land-use and activity systems, on models of supply design and on the analysis of within-day dynamic systems.

Transportation systems theory would, however, be of little use for practical problems without a set of methodologies operationalizing it. The latter allow the construction of systems of mathematical models, which are consistent with the theory and able to simulate the relevant elements of different transportation systems in the real world. These methodologies range from the rules for defining a supply network model to the techniques for estimating travel demand to algorithms for the solution of large-scale problems. Transportation system methodologies use the results of several disciplinary areas and, taken as a whole, make up the technical resources of transportation systems engineers and analysts.

This book attempts to address both general theory and practical methods, and should be useful to readers with different needs and backgrounds. The various topics are presented, wherever possible, with a gradually increasing level of detail and complexity. It includes a series of topics, which can be used as the basis for graduate and post-graduate courses on transportation system engineering, as well as other fields, e.g. economics and regional sciences. Some sections deal with topics of interest for specific applications or still at a research level; the exclusion of these topics, marked with an asterisk, should not detract from the understanding of later chapters. The required mathematical background includes calculus, numerical analysis, optimization techniques, graph and network theory, probability theory and statistics. The general structure of the book is as follows.

Chapter 1 defines a transportation system, identifies its components and the assumptions on which the theory and the models described in later chapters are developed.

Chapters 2 to 5 explore the theory of transportation systems under the traditional assumption of intra-periodic stationarity of the relevant variables. In particular, Chapter 2 introduces supply models and describes the networks representing transportation services, formalizes supply models, the general relationships between flow and performance variables and gives some examples of the models that can be used to represent different supply systems. The appendix to this chapter reviews the main results of traffic flow theory, and queuing theory needed to develop link performance functions. Chapter 3 describes the theoretical basis and the mathematical properties of random utility models; these are the tools most widely adopted to simulate travel behavior of transportation system users. Chapter 4 describes mathematical models that simulate the different aspects of passengers and freight transportation demand, introduces their theoretical formulations and provides several examples.

Chapter 5 defines traffic assignment models, which simulate the interactions between transportation demand and supply, and studies their theoretical properties . Assignment models simulate the interactions between transportation demand and supply. Most of the chapter concerns network equilibrium models, both deterministic and stochastic, with rigid and elastic demand and single or multiple user classes. Some references are also made to recent inter-period (day-to-day) dynamic modeling approaches including both deterministic and stochastic process models.

Chapter 6 extends the results of previous chapters to intra-period (within-day) dynamic systems. In particular, this chapter addresses extensions of supply, demand and supply-demand interaction models to intra-period dynamic systems, both for continuous and scheduled services.

Chapters 7, 8, 9 and 10 discuss methodological aspects related to the applications of transportation system engineering. Chapter 7 describes the algorithms commonly used to efficiently compute network flows resulting from the intra-period static assignment models described.

Chapter 8 explores different methods for estimating transportation demand. Methodologies derived from statistics and econometrics are applied to the estimation of present transportation demand in a given area on the basis of sampling surveys and to the specification and calibration of demand models for the simulation of demand. The chapter also discusses the techniques that can be used to estimate present demand flows and model parameters by using aggregate information, specifically traffic flow counts. Chapter 9 briefly describes several supply design models and algorithms. These can be applied to set the values of unknown parameters defining the design problem at hand by optimizing different types of objective functions under various constraints. Design problems related to network topology, performance parameters and pricing are introduced with respect to both road and transit networks.

Finally, Chapter 10 describes the fields of application of transportation systems engineering, the decision-making process in transportation systems and the role of quantitative methods in such a process. The chapter also briefly introduces some common project evaluation methods, namely Cost-Benefit and Multi-Criteria analyses.

The book also includes an appendix which, to facilitate reading, summarize the main results of some basic disciplines (numerical analysis and optimization theory, as well as relevant algorithms), used in the previous chapters.

Different reading paths can be followed in relation to different theoretical interests; for example, a path focusing on demand analysis could consist of Chapters 3, 4 and 8, while one focusing on transportation systems design and planning could consist of Chapters 2, 5, 7, 9 and 10.

An effort has been made throughout the book to give credit to the proper authors quoted from the literature. Nevertheless, for well-known results credits may be unintentionally omitted or misplaced. Apologies are submitted in advance for any error of this type.

A book of this scope and magnitude cannot be completed without the help and the assistance of several individuals.

Giulio Erberto Cantarella shared the whole process leading to the structure of the book and the choice of its contents. He also contributed directly co-authoring Chapters 5 and 7 as well as Appendix 2A.

Almost all topics covered in this book were discussed over the years with Agostino Nuzzolo, who also co-authored the section 6.5 on dynamic traffic assignment for scheduled services.

Andrea Papola revised the Italian version of the book with intelligence and enthusiasm, making many valuable suggestions especially in Chapters 3 and 4.

Mariano Gallo has contributed to the research and preparation of Chapters 2 and 9, Pierluigi Coppola contributed to Chapter 6 and Alessandra Improta to Chapter 8.

Francesca Pagliara worked on the preparation of Chapter 10 and coordinated the whole process of translation into English.

Raimondo Polidoro supported the preparation of Chapters 5 and 7, and worked hard on the final lay-out of the book and is to be credited with the calculations and the drawing of most figures and tables.

Alan Erera and Karen Smilowitz read the proofs of the whole book making several valuable comments on the text and the English presentation.

Last, but not least, Silvana Carro patiently and professionally typed the whole text through its many versions.

The Author wishes to express his sincere gratitude to all these persons even though, as is obvious, he remains the only responsible for any mistakes.

1 TRANSPORTATION SYSTEMS

1.1. Definition

A *transportation system* can be defined as the combination of elements and their interactions, which produce the demand for travel within a given area and the supply of transportation services to satisfy this demand. This definition is general and flexible enough to be applied to different contexts. The specific structure of the system is defined by the problem itself (or class of problems) for whose solution it is employed.

Almost all of the components of a social and economic system in a given geographical area interact with different levels of intensity. However, it is practically impossible to take into account every interacting element to solve a transportation engineering problem. The typical system engineering approach is to isolate those elements, which are most relevant to the problem. These elements, and the relationships among them, make up the *analysis system*. The remaining elements belong to the *external environment* and are taken into account only in terms of their interactions with the analysis system. As will be seen later, transportation system engineering is oriented to the design and evaluation of transportation supply projects. In general, the analysis system includes the elements and the interactions that are expected to be significantly affected by the projects under consideration. It follows that there is a strict *interdependence between the identification of the analysis system and the problem to be solved.* The transportation system of a given area can also be seen as a sub-system of a wider territorial system with which it strongly interacts. The extent to which these interactions are included in the analysis system, or else in the external environment, depends on the specific problem.

These concepts can be clarified by some examples. Consider a city (or an urban system) consisting of a set of households, workplaces, services, transportation facilities, government organizations, regulations, etc. Within this system, several sub-systems can be identified, including the activity and transportation systems both relevant for our purposes (see Fig. 1.1.1).

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

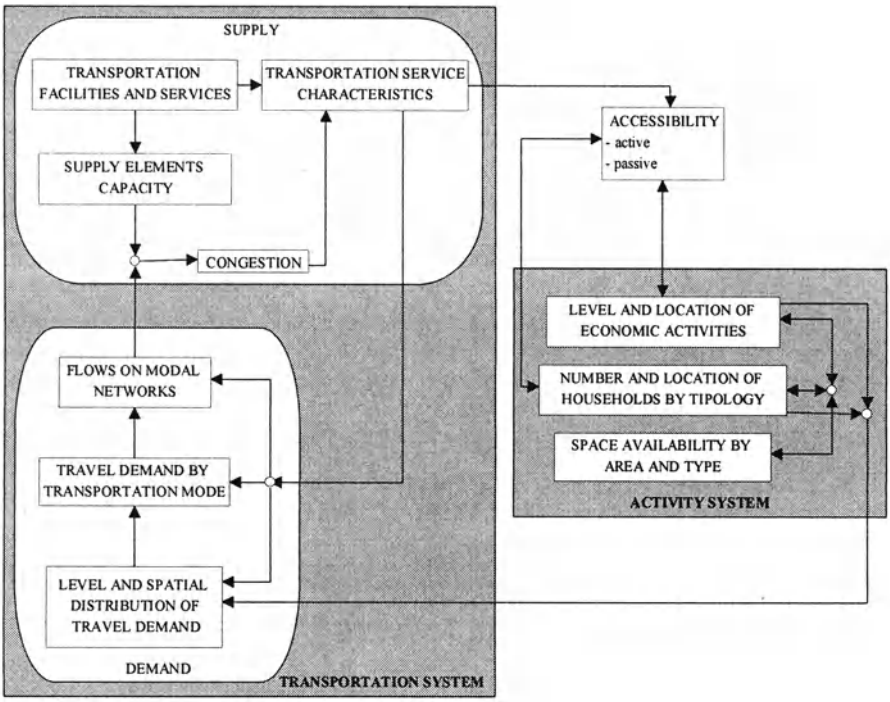


Fig. 1.1.1 Relationships between the transportation system and the activity system

The *activity system* of an urban area can be schematically decomposed into three sub-systems consisting of:

- the households divided into categories (by income level, life-cycle, composition, etc.) living in each zone;
- the economic activities located in each zone and divided by sectors (different industrial and service sectors), by economic (e.g. added value) and physical (e.g. the number of employees) indicators;
- the floor-space (or volumes) available in each zone for various uses (industrial production, offices, residences, shops, building areas, etc.) and relative market prices (real estate system).

The different components of the activity system interact in many ways. For example, the number and typology of the households living in the various zones depend on employment opportunities and their distribution, and therefore on the sub-system of economic activities. Furthermore, the location of some types of economic activities (retail, social services such as education and welfare, etc.)

depends on the distribution of the households. Finally, the households and the economic activities in each zone depend on the availability of specific types of floor-space (houses, shops, etc.) and on the relative prices. A detailed analysis of the mechanisms underlying each sub-system of the activity system is beyond the scope of this book. However it should be noted that the relative "accessibility" of the different zones, ensured by the transportation system, is extremely relevant to many of these mechanisms. The *transportation system* can be split into two main components: demand and supply.

The distribution of households and activities in the area is the determinant of *transportation demand* deriving from the need to use different urban functions in different places. Household members are the users of the transportation supply system and make "mobility choices" (holding a driving license, number of cars, etc.) and "travel choices" (trip frequency, time, destination, mode, path, etc.) in order to undertake activities (work, study, shopping, etc.) in different locations. The result of these choices is the transportation demand; i.e. the number of trips made among the different zones of the city, for different purposes, in different periods, by means of the different available transportation modes. Similarly, economic activities transport goods that are consumed by the households or by other economic activities. Goods movements make up the freight transportation demand.

Both mobility and travel choices are influenced by some characteristics of the transportation services offered by the different travel modes (individual car, transit, walking). These characteristics are known as level-of-service or performance attributes and include travel times, monetary costs, service reliability, riding comfort, etc. Thus, the choice of destination may be influenced by the travel time and cost needed to reach each destination. The choice of departure time depends on the travel time to the destination. The choice of transportation mode is influenced by times, costs, reliability of the available modes.

The characteristics of transportation services depend on the *transportation supply*, i.e., the set of facilities (roads, parking spaces, railway lines, etc.), services (transit lines and timetables), regulations (road circulation and parking regulations), and prices (transit fares, parking prices, road tolls, etc.) producing travel opportunities. The physical elements of the transportation supply system have a finite capacity; i.e. a maximum number of users that can be served in a given time interval.

Individual trips can be aggregated into users flows, i.e. the number of users on the physical elements of the supply system in a given time interval. Examples are automobile and truck flows on road sections, passenger flows on transit lines, and so on.

When flow approaches capacity, the interactions among users increase and *congestion* effects are triggered. Congestion can significantly deteriorate the performances of transportation services for the users, e.g. travel times, service delays, fuel consumptions all increase with congestion. Congestion can also have other "external" negative effects (such as noise, air pollution and visual impacts in

the case of road traffic). Congestion can have cross-modal effects; e.g. road congestion can influence the performances of surface transit services.

Finally, transportation performances influence the relative accessibility of different zones of the urban area by determing, for a given zone, the "cost" of reaching other zones ("active" accessibility), or being reached from other zones ("passive" accessibility). As has been noted, both these accessibilities influence the location of households and economic activities and ultimately the real estate market. For example, in choosing the residence zone, households take into account the active accessibility to the workplace and to other services (commerce, education, etc.); the location of economic activities is chosen taking into account the passive accessibility from its potential clients; the location of public services should be chosen taking into account the passive accessibility from the users, and so on.

An urban transportation system contains many feedback cycles, i.e. cycles of mutual interdependence between the various elements and sub-systems, as shown in Fig. 1.1.1. The innermost cycle, i.e. the one involving the least number of elements and which usually has a shorter "reaction time" to any perturbation, is the interaction between flows, congestion and costs on modal networks, and on the road network in particular. The trips between the various zones made with a given mode (e.g. the car) use different paths and result in traffic flows on the different supply elements (e.g. road sections). Because of congestion, these flows influence travel times and other characteristics of the different paths which, in turn, influence path choices.

There are outer cycles, i.e. cycles influencing several choice dimensions whose changes occur over longer time periods. These cycles involve the distribution of trips among the possible destinations and the alternative modes. Modal origin-destination demand flows induce traffic flows that, due to congestion, modify the service characteristics, which in turn influence destination and mode choices.

Finally, there are other cycles that span even longer time periods in which activity locations and transportation demand interact. Again, through user flows and congestion, travel demand influences the accessibility of the different zones of the urban area and therefore the location choices of households and firms.

The aim of transportation systems engineering, as will be seen in greater detail in Chapter 10, is to design transportation supply projects by using the quantitative methods described in the following chapters. The projects may have very different "dimensions" and impacts, and consequently the boundaries of the analysis system and the external environment will be different.

If the problem at hand is the long-term planning of the whole urban transportation system, including the construction of new motorways, railway lines, parking facilities, etc., the analysis system has to include the entire multi-modal transportation system and possibly its relationships with the urban activity system. In fact, the modifications in transportation performance implied and the time needed to implement the project are such that all the components of transportation and activity systems will likely be affected.

There are cases, however, in which the problem is more limited. If, for example, the aim is to design the service characteristics of an urban transit system without

building new infrastructures (and without implementing new car restriction policies), it is common practice to include in the analysis system only the elements related to public transportation (demand, services, prices, vehicles, etc.). The rest of the transportation system is included in the external environment interacting with the public transportation system.

As will be seen in the following chapters, the above examples can easily be generalized to areas of different size (a region, a whole country, etc.) and extended to the case of freight transport.

Transportation systems are generally described as *complex systems*, i.e. systems made up of several elements with non-linear interactions and several *feedback cycles*. Furthermore, the unpredictability of most features of the system, such as the travel time needed to cover a road section or the users' choices, would require the state of such a system to be represented by random variables. These random variables are often substituted by their expected values as a first approximation.

Transportation systems engineering, and its quantitative methods, focus on the analysis and the simulation of the elements and the relationships that make up the transportation system, considering the activity system as exogenously given. More specifically, the influence of the activity system on the transportation system, and in particular on travel demand, is considered, while the inverse influence of accessibility on activity location and level is usually neglected. However, this conventional demarcation is rapidly vanishing and the whole activity-transportation system is studied more and more often in transportation system projects, though with different levels of detail with respect to other disciplines, such as regional sciences and spatial economics.

In the following sections of this chapter, transportation supply and demand systems will be described and characterized in more detail, introducing the general framework and the basic assumptions used in the theory and models described in this book.

1.2. Transportation system identification

The identification of the transportation system, i.e. the definition of the elements and their reciprocal relationships making up the analysis system, is schematically carried out in three phases:

- identification of relevant spatial and supply characteristics;
- definition of relevant components of transport demand;
- identification of relevant temporal dimensions.

Some comments on the different phases will be given below, it should be stated in advance that system identification cannot be reduced to the mere application of a set of rigid rules. On the contrary, it is the result of a combination of theory and experience, usually referred to as "professional expertise".

1.2.1. Relevant spatial and supply characteristics

The identification of relevant spatial and supply characteristics consists of three phases:

a) delimitation of the study area;
b) subdivision of the area into traffic zones (zoning);
c) definition of the relevant infrastructures and services.

The first two phases are preliminary to the building of a demand/supply model since they define the spatial extension of the system, its level of spatial aggregation. The next phase relates to the identification of supply characteristics systems, which is strongly related to zoning.

Study area

This phase defines the geographical area including the transportation system under analysis and most of the project effects.

First, the analyst must consider the decision-making context and the type of relevant trips: commuting, leisure, etc. Generally, most trips would have origin and destination inside the study area. On the other hand, the study area should include possible alternative for re-routing for destination, and so on.

The limit of the study area is usually known as the *area cordon* or *boundary*. Outside this boundary is the external area, which is only considered by its connections with the analysis system. For instance, the study area might be a whole country if the transportation project is at a national level, a specific urban area, or part of an urban area for a traffic management project.

Zoning

In principle, the trips undertaken in a given area may start and end at any point. To model the system, it is necessary to subdivide the area (and possibly portions of the external area) into a finite number of discrete *traffic zones*. Trips between two different zones are known as *inter-zonal* trips, while *intra-zonal* trips are those starting and ending within the same traffic zone (generally not loaded on the network).

Traffic zones may consist of entire cities or groups of cities in regional or national wide projects, or of a few blocks in urban traffic projects. Thus zoning a given area implies the approximation of the actual starting and terminal points of interzonal trips with single points (*zone centroids*). Zoning is related to the subsequent phase of selection of the relevant supply elements. A denser set of elements usually corresponds to a finer zoning, i.e. a larger number of traffic zones, and vice versa (see Fig. 1.2.1). For example, if the urban system includes public transport, it is common practice to consider smaller traffic zones than for a system including individual modes. This is due to the need to simulate realistically walking access to stops and/or stations through the distance from the zone centroid.

There are several possible zoning systems of the same study area and for the same problem. However, some general "zoning rules" are often followed. Physical geographic separators (e.g. rivers, railway lines) are conventionally used as zone boundaries since they prevent "diffuse" connections between adjacent areas and therefore usually imply different access conditions to transportation facilities and services. Traffic zones are often obtained as aggregations of administrative areas (e.g. census sections, municipalities or provinces). This allows to associate to each zone the relevant statistical data (population, employment, etc.) usually available for such areas. A different level of zoning detail can be adopted for different parts of the study area depending on the precision needed. For example, smaller zones may be used in the neighborhood of a specific element (e.g. a new road, railway, etc.) for which traffic flows and impacts must be simulated more precisely.

Traffic zones should aggregate parts of the study area which are "homogeneous" with respect both to their land-use (e.g. residential or commercial zones in urban areas or rural municipalities in extra-urban areas) and to their accessibility to transportation facilities and services.

A larger number of zones provide a more precise representation of the real system and a lower incidence of intra-zonal trips, whose effects on the physical network cannot be simulated. On the other hand, a large number of zones increases the computational burden of any model. In practice, achieving a reasonable compromise between these two conflicting requirements depends once again on the particular type of project.

Centroid nodes are fictitious nodes representing the actual starting and terminal points of trips beginning or ending in each traffic zone. In principle, several centroids could be located within a single zone to represent the terminal points of different trip types, e.g. one centroid may represent origins and another destinations. In practice, however, a single centroid is usually associated with each zone.

The external area is usually subdivided into larger traffic zones. In fact, external zones are only used to characterize those trips using, at least partially, the transport system within the study area. External zones are also represented by zone centroids; such centroids can be located at boundary points (road sections, stations, ports, airports, etc.) through which the trips from and towards the external zones enter and leave the study area. Alternatively they can be located baricentrically with respect to the activity systems of external zones and connected to a network of facilities and services used to reach/leave the study area.

Relevant infrastructures and services

Relevant infrastructures and/or services are identified on the basis of their role in connecting the traffic zones in the study area and the external zones. This generates a close interdependence between supply selection and zoning.

Since the flows on network elements resulting from intra-zonal trips are not simulated, a very fine zoning with a coarse base network will probably cause overestimation of the traffic flows on the included network elements.

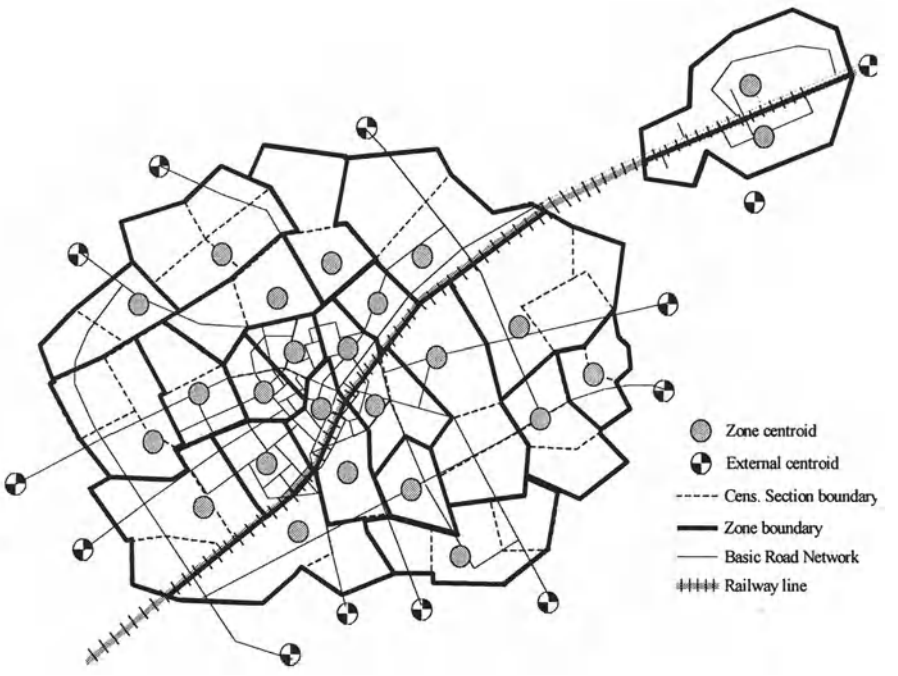


Fig. 1.2.1 Zoning and base network

Vice versa, a very detailed base network with a coarse zoning may cause underestimation of some traffic flows.

Infrastructures and services may relate to only one transport mode (e.g. road, railway or air services) or to several modes. The former will be referred to as a mono-modal system, the latter as a multi-modal system.

The set of elements considered for a given application is sometimes known as "*base network*" and is usually represented graphically by highlighting the infrastructures used by the selected transportation services. The functional characteristics needed to build the mathematical model (network) of the transportation supply are often associated to the selected facilities. For example, in urban road systems, the road sections and their main traffic regulations such as one-way, no turn, etc. are indicated (see Fig 1.2.2). For scheduled service systems, the infrastructures over which the service is operated (road sections, railways, etc.) will be indicated, together with the main stops or stations, the lines traveling along the physical sections, etc.

The identification of the relevant elements is obviously easier when all the services and/or the infrastructures play a role in connecting traffic zones, as may be the case for a national airways network. In the case of road networks, only a subset is relevant in connecting the different zones. In urban areas, for example, local roads are usually excluded from the base network of the whole area, while they may be included in the base networks of spatially-limited sub-systems (a neighborhood or part of it). Similarly, when dealing with a whole region, most roads within each city will not be included in the base network.

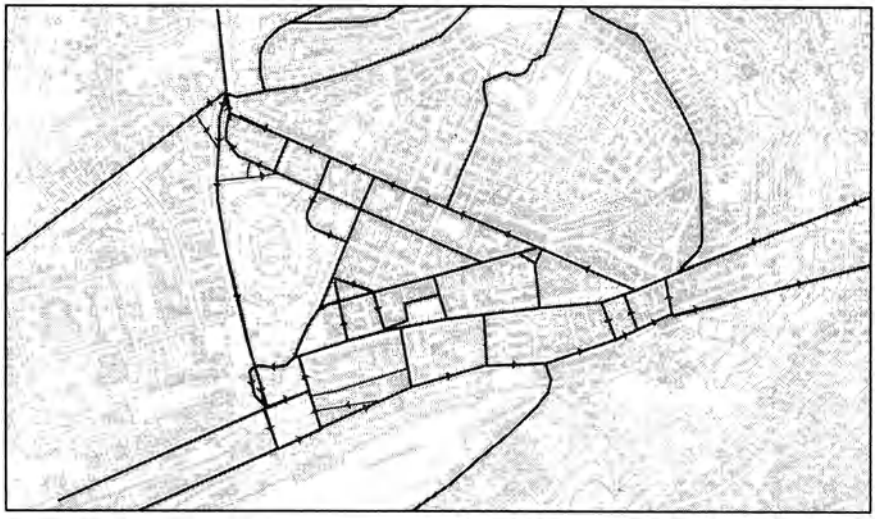


Fig. 1.2.2 Base road network for a portion of urban area.

1.2.2. Relevant components of transport demand

Passengers and goods moving in a given area demand transportation services supplied by the system. This demand plays a central role in the analysis and modeling of transportation systems since transportation projects are motivated by the need to satisfy transportation demand, sometimes modifying some of its characteristics as in travel demand management policies. In turn, traveler choices can significantly affect the performance of supply elements through congestion (see Chapters 2 and 5).

Except for a limited number of cases, travel does not provide "utility" in itself, but is rather an auxiliary activity for other activities carried out in different locations. Travelers make work-, school-, and shopping-related trips. Goods are shipped from production sites to markets, warehouses or industrial plants to be further processed, and so on. Transport demand in economic terms is therefore a "derived" demand, the result of the interactions between the activity system of the area and the transport

services and facilities, as was seen in section 1.1, as well as of the habits underlying travel behavior.

A *transport* or *travel demand flow* can formally be defined as *the number of users with given characteristics consuming the services offered by a transport system in a given time period.* It is clear that transport demand flows result from the aggregation of individual trips or shipments made in the study area during the reference period. A *trip* is defined as the act of moving from one place (origin) to another (destination) using one or several means or modes of transport, in order to carry out one or more activities. A sequence of trips, which follow each other in such a way that the destination of the previous trip coincides with the origin of the next, will be defined as a *journey* or *trip-chain.* With passenger travel, journeys usually start and end at home; for example, a journey home-workplace-shopping area-home consists of three trips. The users of a system, or the trips they undertake, can be characterized with reference to several factors, as described below.

Spatial characterization is important because of the very definition of travel. Trips can be subdivided by place (zone or centroid) of *origin* and of *destination* and demand flows can be arranged in origin-destination matrices (*O-D matrices*). These matrices (see Fig. 1.2.3) have a number of rows and columns equal to the number of zones; the generic entry d_{od} gives the number of trips (or shipments) made in the reference period from origin zone o to destination zone d (*O-D flow*).
Some aggregations of the elements of the O-D matrix are also useful. The sum of the elements of the row o:

$$d_{o.} = \sum_d d_{od}$$

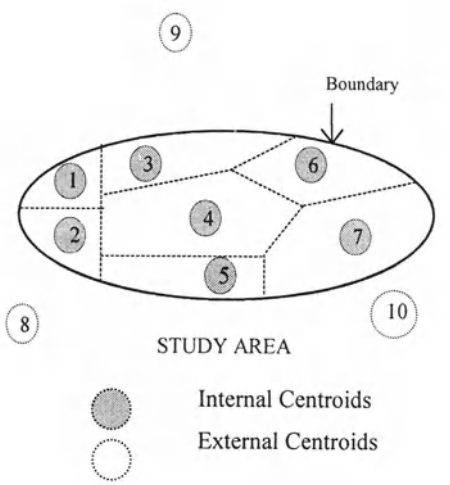
accumulates the total number of trips "starting" from the generic zone in the reference period and is known as the *flow "emitted" or "generated" by zone o.* The sum of the elements of the column d accumulates the number of trips arriving in zone d in the reference period:

$$d_{.d} = \sum_o d_{od}$$

and is known as the flow "*attracted*" by zone d. The total number of trips carried out in the study area in the reference interval is indicated by $d..$:

$$d.. = \sum_o \sum_d d_{od}$$

In *exchange trips*, the origin and the destination are one within and the other outside the study area.



O-D MATRIX

Destination zones

Origin zones	1	2	3	4	5	6	7	8	9	10
1	•									
2		•								
3			•	Internal trips				Internal-External Exchange trips (I-E)		
4				•						
5					•					
6						•				
7							•			
8	External-Internal Exchange trips (E-I)							Crossing trips (E-E)		
9										
10										

• = intrazonal trips

Fig. 1.2.3 Trip types and their identification in the Origin-Destination matrix.

Finally, *crossing trips* have both the origin and the destination external to the study area, but transverse the study area, i.e. use the transportation system under study. Fig. 1.2.3 is a schematic representation of the three types of trips and their position in the O-D matrix.

Transport demand can also be classified by *user* and *trips characteristics* relevant to the analysis. In general, the characteristics of the users are distinguished from those of the particular trip undertaken. The former, such as income group or driving-license holding, are usually defined socio-economic characteristics. Groups of users who are homogeneous with respect to the socio-economic characteristics relevant to the specific problem are also referred to as *market segments*. In the study of different pricing policies, for example, market segments can be defined with respect to income. In the case of goods, the "user's" characteristics can be those of the shipping firm (such as dimensional class, type of plant, production cycle and so on).

Trip characteristics relate to the particular trip. Typically a pair of purposes, or activities, can be associated with each trip. In the case of passengers they can be the purpose for traveling from (activity carried out at) the origin and the purpose for traveling to (activity carried out at) the destination, such as home-work trips, work-shopping trips, and so on. Furthermore, a whole sequence of purposes (activities) can be associated with a journey or trip-chain.

Other trip characteristics may include desired arrival or departure times, mode, etc. for passenger travel, or consignment size, type of goods (economic sector, perishability, value, etc.) for freight transport.

1.2.3. Relevant temporal dimensions

A transportation system operates and evolves over time, and both travel demand and supply characteristics generally vary in different time intervals. For example, the number of trips undertaken in an urban area or the frequencies of transit schedules vary at different times of the day, on different days of the week and so on.

While space has always been recognized as a fundamental dimension of transportation systems, time has traditionally been overlooked. The determination of the time intervals relevant for analysis and simulation as well as the assumptions on the system variability within those intervals are crucial and depend once more on the purpose of the analysis. Two different time dimensions are usually relevant for design and evaluation of transportation projects. The design of the elements of a transportation system (e.g. a road cross-section, traffic lights at an intersection or the frequency of a transit line) usually requires information related to short *maximum-load periods* (such as the rush hour or part of it). On the other hand, economic and financial evaluations of a transport project require information over longer periods, comparable to the "technical life" of the project.

In general two significant time intervals can be defined. The *analysis interval* (*period*) is the period of time relevant to study a given system (both in the past and over a hypothetical future horizon)[1]. A *reference* or *simulation interval* (*period*) is a

period of time for which the system is simulated (using the mathematical models described later in this book). The analysis interval is usually longer and may include several simulation intervals. For certain applications the analysis period may span several years, but the system is simulated only for a limited number of simulation intervals, (say one average day per year), and the results obtained are expanded to the whole analysis period. On the other hand, some applications require only the simulation over a single reference period (e.g. the a.m. peak period) on an average weekday.

With respect to the system dynamics within the reference interval, two hypotheses can be made corresponding to two different modeling approaches. Mathematical models of transport systems are traditionally built on the assumption of *intra-period stationarity*. It is assumed that demand and supply remain constant over a period of time long enough to allow the system to reach a *stationary or steady-state condition*. During that period all the relevant characteristics, such as demand, traffic flows and supply performances are constant on average and independent of the particular instant at which they are measured. The other approach is based on the hypothesis of *intra-period dynamics*, i.e. the variations of demand and/or supply within the reference interval are explicitly taken into account and simulated. It should be noted that in practice also intra-period dynamic models assume that some elements of the system, e.g. activity-system variables or global travel demand, are constant within the simulation interval.

If both demand and supply remained (approximately) constant over the whole analysis interval, any sub-interval of the analysis interval could be adopted as the reference period. The results obtained for one of such intervals could be extrapolated to the whole analysis period. However, the assumptions made in the identification of the simulation period (i.e., the intra-period stationarity of relevant variables) usually cannot be extended over the whole analysis interval; thus, the latter is typically decomposed into sub-intervals, corresponding to different reference intervals[2].

The temporal variations of the system characteristics within the analysis interval, theoretically can be decomposed into three classes corresponding to the decomposition of the time series of the relevant variables.

a) *Long-term variations* or *trends* are the global-level and/or systematic variations that can be identified by averaging over several reference periods. For example, if reference intervals are single days, the trend consists of variations in the total level and/or in the structure of the average annual demand, observed over several years. In this case, the daily demand is averaged over 365 elementary periods. Long-period variations are often the result of structural changes in the socio-economic variables underlying transport demand, or in transport supply. For example, variations in the level of economic activity, production technologies, available income, individual vehicle ownership, socio-demographic characteristics of the population, life-styles, urban migration, and in the stock of transportation facilities and services have significantly modified

the level and the structure of passenger and freight transport demand over the years (see Fig. 1.2.4).

b) *Cyclic (seasonal) variations* occur within the analysis interval and involve several reference periods. These variations repeat themselves cyclically and can be observed by averaging over a number of cycles. This is the case, for example, with variations of the daily demand for different days of the week or with variations at various times within a typical day. Figure 1.2.5 shows the breakdown of daily demand by trip purpose against the time-of-day in an urban area. In an analysis interval, several cyclic variations with different cycle lengths may occur and overlap with long term variations. For example, in an urban transportation system, demand and supply change over an analysis period of several years (long-term variation), but they also vary cyclically over the different months of the year, the days of the week and the hours of each day.

c) *Inter-period variations* are the variations in demand and supply over (reference) periods with identical characteristics once that trend and cyclic variations have been accounted for. This is the case with variations in the demand during a.m. peak hours of different days with similar characteristics. These fluctuations cannot be associated with systematic events, i.e. variations in the input variables taken into account in the model representing the system. Transport demand results from the choices made by a large number of users; its actual value in a period (e.g. a day or part of a day) depends both on the unpredictable behavioral elements connected with these choices and on the influence of the choices made in previous periods. Similarly, the actual values of some key supply parameters, such as road capacities or travel times for a given day, may vary due to unpredictable events, e.g. an accident. Variations in demand and supply between successive reference periods, e.g. hours within typical days, are usually known as *inter-period dynamics.*

Figure 1.2.6 shows the traffic flows counted over some road links in a sequence of successive reference periods.

As already mentioned, in real systems the three types of dynamics overlap and their identification depends to a great extent on the perspective adopted. Some models can simulate endogenously variations of some relevant parameters within a typical day, which in this case may be assumed as the simulation period. Other models may require different exogenous input variables to model variations over different hourly periods of the day; in this case, single hours may be the best simulation periods. Moreover, different application contexts usually require different assumptions on the relevant temporal dimensions.

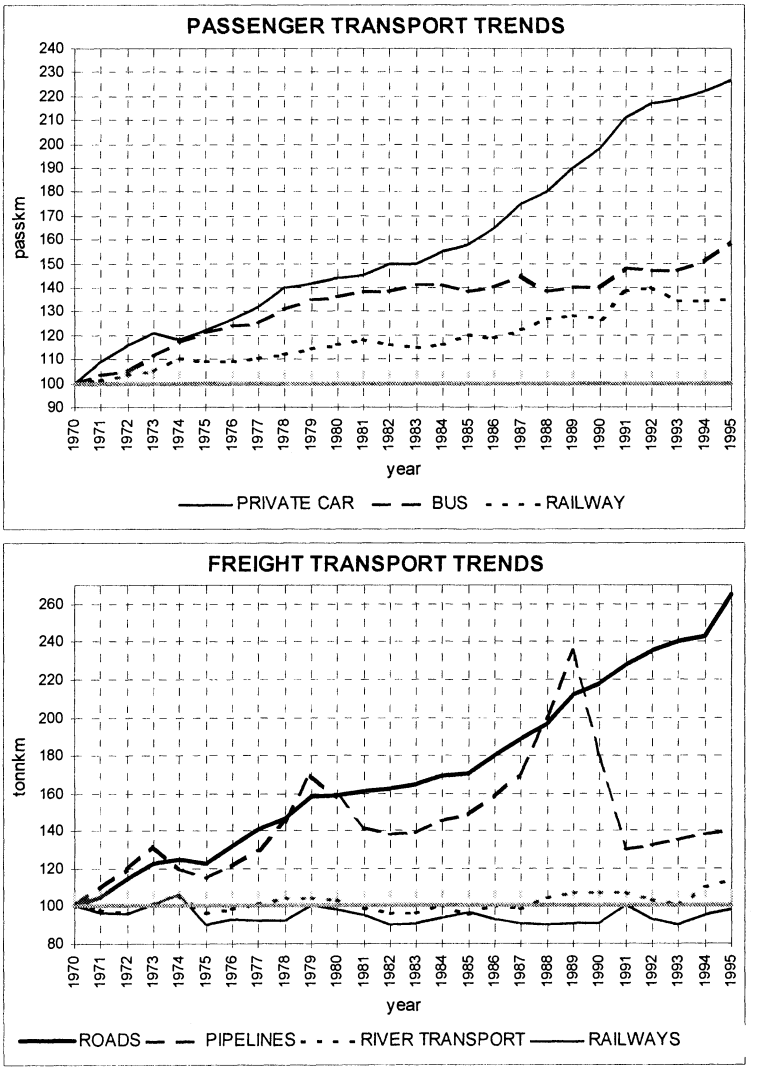


Fig. 1.2.4 Long-period trends of passenger and freight demand: average European values.

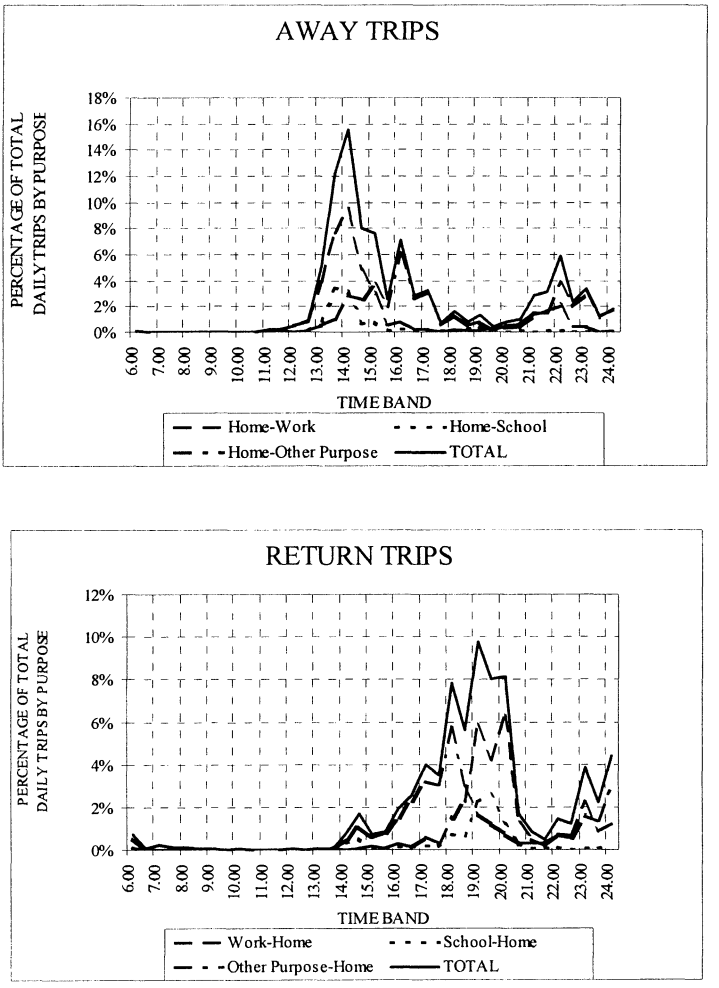


Fig. 1.2.5 Breakdown of urban travel demand by time of the day and purpose

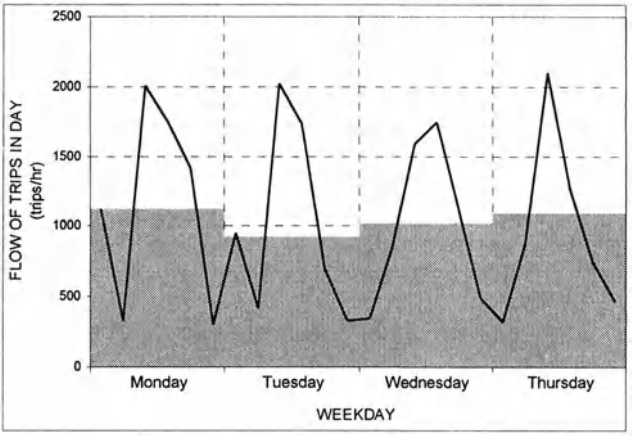


Fig. 1.2.6 Road traffic flows on successive weekdays.

For example, a freight system project may require an analysis period several years long; however, no significant congestion is expected. In this case might be appropriate to consider long-term variations of the system over successive years, and account for seasonal variations by considering some typical months as simulation periods while ignoring cyclic variations within the month.

For a project with a short-term horizon, such as the traffic plan of an urban area, the long-term trend of daily demand (say over several years) can be ignored. The analysis period could consist of one or more typical days (e.g. average week and weekend days). Cyclic variations may be modeled as hourly variations within the typical day. Simulation periods may encompass the a.m. and p.m. peak and off-peak hours during which the system is assumed to be stationary. Alternatively, the analyst may consider a different perspective where the analysis period is an entire week, cyclic variations are relative to both days of the week and hours of the day, reference periods encompass full days. In this case the models explicitly simulate the distribution of demand and supply performances over sub-intervals of each day following an intra-period dynamic approach (see Figure 1.2.7).

In conclusion, the main assumptions regarding determination of the temporal dimensions of a particular study include:

- determination of the analysis interval and how to model long-term trends of exogenous variables;
- determination of reference (or simulation) intervals considered relevant to account for the cyclic variations of transport demand and supply;
- assumption on the variability of relevant system parameters within each selected reference period (intra-period dynamic or static models);
- inference of relevant information on the system by combining the results relative to each simulation interval.

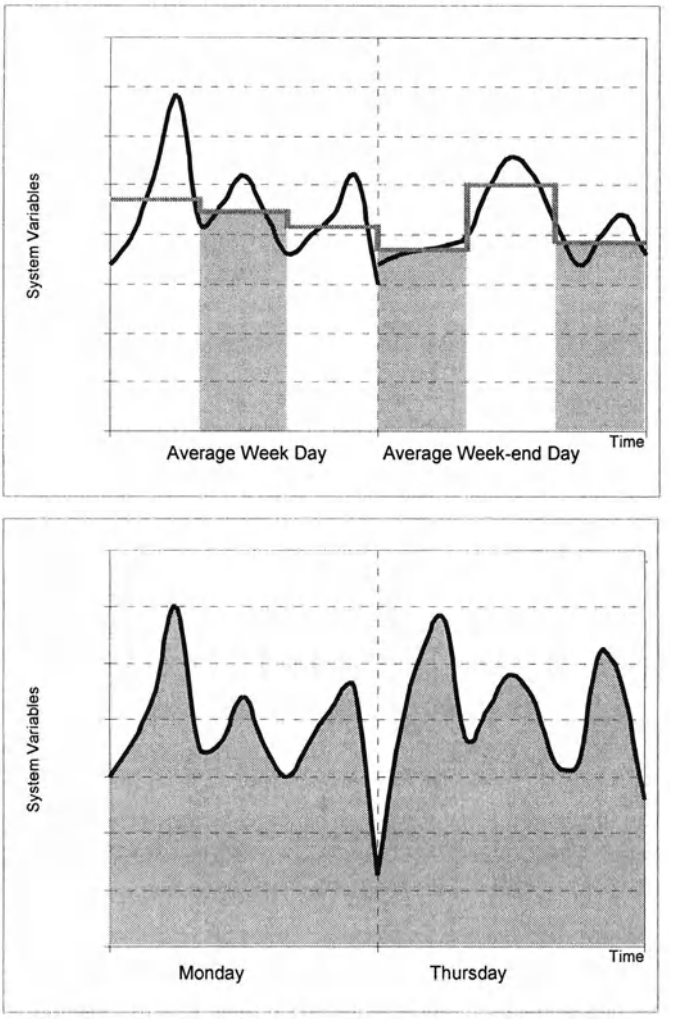


Fig. 1.2.7 Alternative reference periods.

1.3. Modeling transportation systems

The relevant interactions among the various elements of a transportation system can be simulated with the mathematical models that will be described in the following chapters. It is useful to anticipate an overview of the various classes of models, which make up the "system" of models simulating a given transportation system. The models and their relationships are described in Fig. 1.3.1.

Supply models, described in Chapter 2, simulate the transportation services available among the different zones with flow network models. More specifically, supply models simulate the performance of transportation infrastructures and services for the users, as well as the main external effects of transport (pollution, energy consumption, accidents). The level-of-service attributes, such as travel time and cost, will be input variables for the demand models. To simulate the performance of single elements (facilities) and the effects of congestion, especially for road systems, supply models use the results of traffic flow theory, which is briefly described in the appendix of Chapter 2.

Demand models simulate the relevant aspects of travel demand as a function of the activity system and of the supply performances. Typically, the characteristics of travel demand simulated include the number of trips in the reference period (demand level) and their distribution among the different zones, the different transport modes, and the different paths.

Other components of travel demand are simulated in specific applications such as the distribution between different time intervals within the reference period. Demand models, which will be described in Chapter 4, can be applied to passenger as well as to freight demand. Travel demand models are usually derived from random utility theory, described in Chapter 3.

Assignment models (or network demand-supply interaction models), studied in Chapters 5 and 6, simulate how O-D demand and path flows load the various elements of the supply system. Assignment models allow the calculation of link flows, i.e. the number of users loading each link of the network representing the transportation supply in the reference period. Furthermore, link flows may affect the transportation supply performances through congestion and therefore may affect the input to demand models. The mutual interdependencies of demand, flows and costs are simulated by assignment models and will be addressed in Chapter 5.

The models described in this book are based on some general assumptions already introduced in the previous sections of this chapter, and summarized below.

a) *Physical and functional delimitation of the system.* The transportation system is assumed to be contained within a defined region (*study area*) and the external area is considered only through its relationships with the analysis system. These relationships are related to both demand (exchange and crossing demand) and supply (transportation infrastructures and services connecting the external area with the analysis system).

b) *Spatial discretization (zoning).* The physical area is subdivided into discrete sub-areas (traffic zones). It is assumed also that the departure and arrival points of all the trips related to a zone are concentrated in an arbitrary point known as *zone centroid.*

c) *Identification of relevant transportation services.* Relevant infrastructures and/or services connecting the internal and external traffic zones are identified.

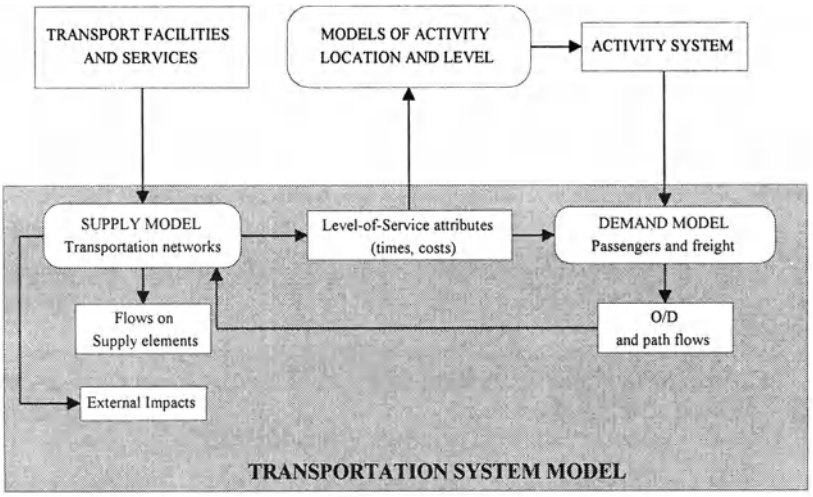


Fig. 1. 3.1 Structure of transportation systems simulation models.

More assumptions about time evolution concern:

i) *Identification of relevant simulation periods.* This refers to the definition of the length of the analysis period, the selection of the significant cyclic variations to be modeled, and finally the identification of the corresponding reference or simulation periods.

ii) *Intra-periodic temporal assumptions.* The intra-periodic stationary approach, adopted in Chapters 2, 4 and 5, assumes that the transport demand with its relevant characteristics and the transport supply have average values constant over a period of time long enough to allow stationary conditions to be reached. In these conditions, the significant variables assume values independent from the reference time. Alternatively, intra-period dynamic models simulate explicitly how supply and some demand dimensions vary within each reference period. Intra-period dynamic models are still at a relatively early stage of development and are discussed in Chapter 6.

iii) *Type of demand-supply interaction.* In the equilibrium approach, it is assumed that the system (after a short time) reaches an *equilibrium configuration* in which demand, flows and costs are mutually consistent. Equilibrium assignment models have been traditionally studied and are described in Chapters 5 and 6. Alternatively, it is possible to adopt an *inter-period dynamic* approach to the modeling of demand-supply interactions by explicitly simulating the evolution of the system over different reference periods. Models of this type are considered in section 5.8.

Finally, traditional transportation models are sometimes integrated with models simulating *activity location* and *production levels*. These models differ according to the size of the study area (urban, regional, and national) and the type of activities considered to be endogenous (i.e. explicitly represented in the model). For example, they may relate to household location in an urban area or to the production level in different sectors of the economy at multi-regional level. Models simulating the transportation system and activity locations are usually referred to as land-use transportation interaction models. This class of models is less widely used than transportation system models, and their systematic analysis is outside the scope of this book. It should be emphasized that most of the concepts underlying land-use transportation interaction models are similar to those discussed in the following. An example of models simulating the interactions between production levels, economic activity location and transportation is described in section 4.6, when looking at freight demand models.

Reference Notes

The definition of a transportation system and its elements can be found in most textbooks covering the analysis and modeling of transportation systems, though with slightly different interpretations. Descriptions of this kind can be found, among others, in Manheim (1979), Sheffi (1985), Ortuzar and Willumsen (1994).

Definitions of transport demand and its characteristics can be found in most textbooks on transportation systems analysis, such as Wilson (1974), Hutchinson (1974), Manheim (1979), Meyer and Miller (1984), Ortuzar and Williumsen (1994).

Descriptive analyses of the structure of transportation demand and its development over time are given in several publications. Examples are the study of the European Conference of Ministries of Transport (ECMT 1992), for passenger transport, and that of the Organization for Cooperation and Economic Development (OCED 1986) for freight transport. Recent overviews of travel demand trends in some transportation markets are in Boyer (1998).

Notes

(1) Obviously it is not possible to forecast the future evolution of a transportation system, with or without the project under study, with absolute confidence. Thus only hypothetical "futures", or scenarios, can be simulated based on a set of assumptions for both exogenous variables and projects on the system. This point is discussed further in later chapters.

(2) Analysis intervals including several stationary sub-periods (e.g. the average day with several homogenous time-bands) could be dealt with by considering a single reference period with average values of the parameters (e.g. travel demand or supply). This possibility, however, could induce severe distortions especially for congested systems (see Chapter 2) as congestion and demand are both highly non-linear phenomena and average flows and performances can significantly differ from flows and performances computed with average characteristics.

2 TRANSPORTATION SUPPLY MODELS

2.1. Introduction

This chapter deals with the mathematical models simulating transportation supply systems. In broad terms a transportation supply model can be defined as a model, or rather a system of models, simulating the performances and the flows resulting from users' demand and the technical and organizational aspects of the physical transportation supply. The general structure of a supply model is depicted in Fig. 2.1.1, where several elements (or sub-models) can be distinguished. The *graph* defines the topology of the connections allowed by the transportation system under study, while the *network loading* or *flow propagation model* defines the relationship among path and link flows. The *link performance model* expresses for each element (link) the relationships between performances, physical and functional characteristics, and flow of users. The *impact model* simulates the main external impacts of the supply system. Finally, the *path performance model* defines the relationship between the performances of single elements (links) and those of a whole trip (path) between any origin-destination pair.

Transportation supply models combine traffic flow theory and network flow theory models. The former ones are used to analyze and simulate the performances of the main supply elements, the latter to represent the topological and functional structure of the system. Throughout this chapter, as stated in Chapter 1, it will be assumed that the transportation system is intra-period (within-day) stationary; extensions of supply models to intra-period dynamic systems will be dealt with in Chapter 6.

The elements composing a transportation supply model will be described in section 2.2 by applying network flows theory to develop an "abstract" supply model (*transportation network*), together with the set of general mathematical relationships between transportation costs and flows on a network (*supply model*).

Successively, some general indications about the applications of network models will be developed in section 2.3. Specific models for transportation systems with *continuous* services (such as road systems) will be described in section 2.3.1; models for *discontinuous* or *scheduled services* (such as bus, train or airplane) will be described in section 2.3.2. Appendix 2A provides a short review of the main results of traffic flow theory.

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

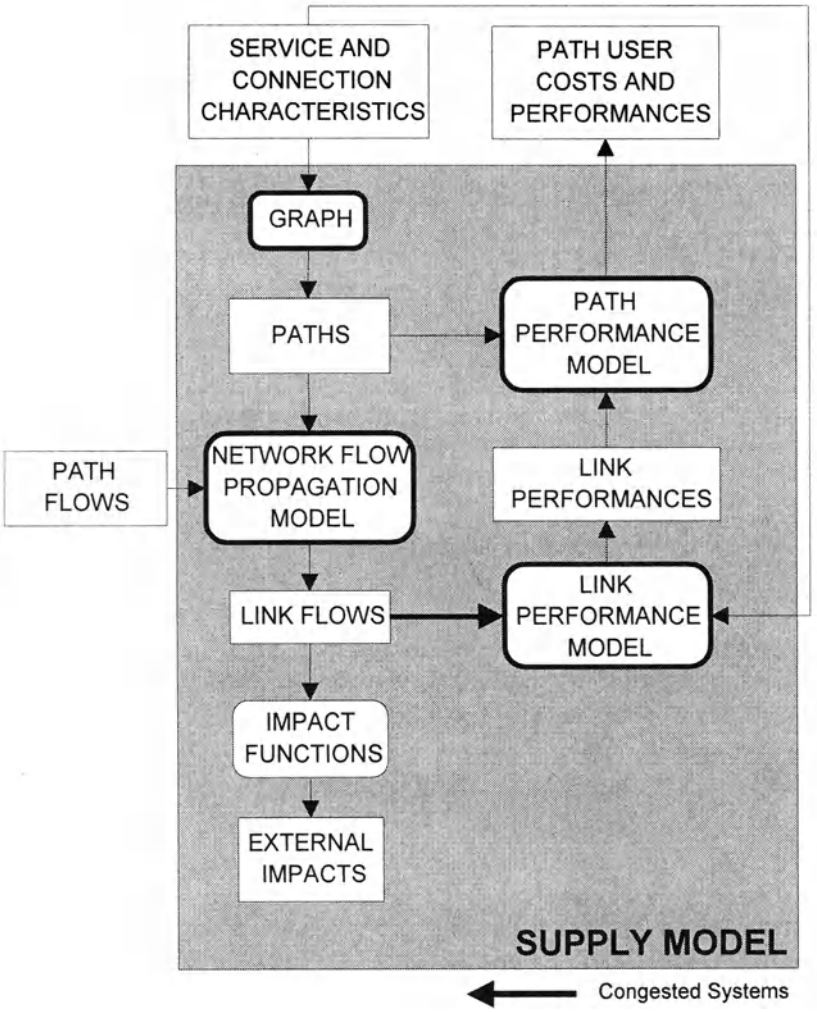


Fig. 2.1.1 Schematic representation of supply models.

2.2. Congested network models

This section provides a general mathematical formulation of transportation supply models, based on congested network flow models. The bases for these models are *graph models*. Successively, *network models*, including link performances and costs, and *network flow models*, including link flows, are introduced. Finally, *congested network (flow) models*, modeling relationships between performances, costs and flows, are developed.

2.2.1. Graph models

A *graph* is defined as an ordered pair of sets: N, the set of elements known as *nodes* or *vertices,* and $L \subseteq N \times N$, a set of pairs of nodes belonging to N, known as *links* or *arcs*. Symbolically, a graph G can be represented by $G = (N,L)$. The graphs used to represent transportation services are generally oriented; i.e., the links have a direction and the node pairs defining them are ordered pairs. A link connecting the node pair (i,j) can also be denoted by a single index, say l, representing its position in the list of all the links of the graph or by the pair of indices, (i,j), relative to the initial and final nodes of the same link (see Fig. 2.2.1).

The *links* in a graph modeling a transportation system represent phases and/or activities of possible trips between different traffic zones. Thus, a link can represent an activity connected to a physical movement (e.g. covering a road) or an activity not connected to a physical movement (such as waiting for a train at a station). Links are chosen in such a way that physical and functional characteristics can be assumed to be homogeneous for the whole link (e.g. the same average speed). In this sense, links can be seen as the partition of trips in segments of equivalent characteristics; the level of detail of such partition can clearly be very different for the same physical system according to the objectives of the analysis.

Nodes correspond to significant events delimiting the trip phases (links). Nodes can correspond to points with different space and/or time coordinates in which the events, represented by the nodes, occur. In *synchronic networks*, nodes are not identified by a specific time coordinate, and the same node represents events occurring at different moments (instants) of time. For example, the different entry or exit times in a road segment, an intersection, a station, may be associated to a single node, representing all the entry/exit events. *Centroid nodes*, introduced in section 1.2.1, represent the beginning and/or the end of individual trips. In *diachronic networks*, on the other hand, nodes may have an explicit time coordinate and therefore represent an event occurring at a given instant. The graphs considered in this chapter are synchronic, since diachronic networks assume a within-period system representation; diachronic graphs for scheduled services will be introduced in Chapter 6.

In a graph representing transportation supply, a *path*, k, is a sequence of consecutive links connecting an initial node (path origin) and a final node (path destination). Thus a path is a sequence of trip phases. Usually, only paths connecting centroid nodes are considered in transportation graphs. These paths are sequences of phases allowing travel from a given origin to a given destination and therefore represent possible trips. On this basis, each path is unambiguously associated with one and only one O-D pair, while several paths can connect the same O-D pair. An example of graph with the different paths connecting the centroid nodes is depicted in Fig. 2.2.1.

A binary matrix called the *link-path incidence matrix*, Δ, can represent the relationship between links and paths. This matrix has a number of rows equal to the number of links, n_L, and a number of columns equal to the number of paths, n_P. The

generic element δ_{lk} of the binary matrix Δ is equal to one if the link l belongs to path k, $l \in k$, and zero, otherwise, $l \notin k$ (see Fig. 2.2.1). The row of the link-path incidence matrix corresponding to the generic link l identifies all the paths including that link (columns k for which $\delta_{lk} = 1$). Moreover the elements of a column corresponding to the generic path k identify all the links that make it up (rows l for which $\delta_{lk} = 1$).

2.2.2. Performance variables and transportation costs

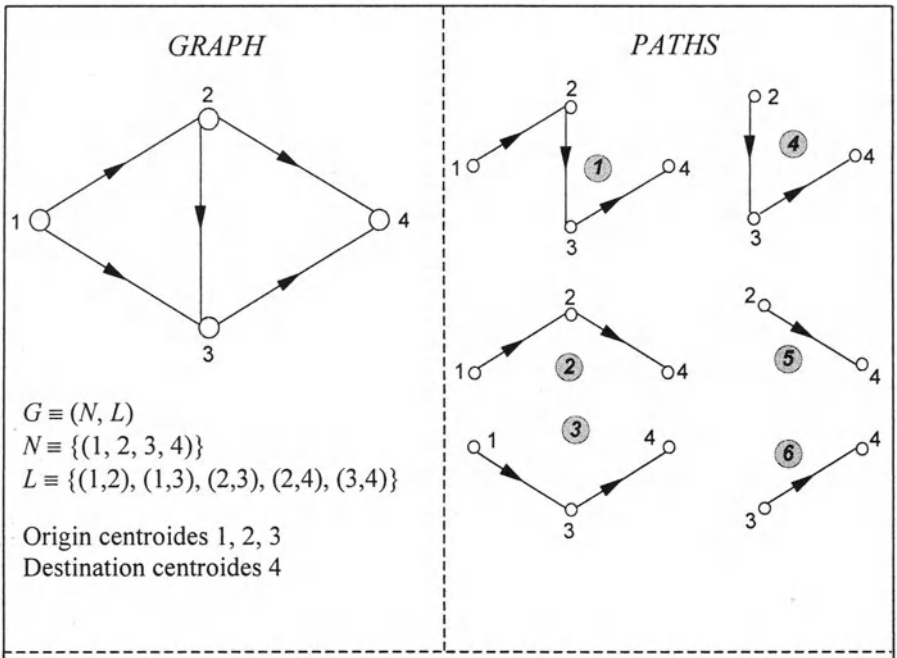
Some variables perceived by users can be associated with individual trip phases. Examples of such variables are travel times (transversal and/or waiting), monetary cost, discomfort, etc. These variables are referred to as *level-of-service* or *performance attributes.* The average value of the n-th performance variable, related to link l, will be denoted by r_{nl}. In general, performance variables correspond to disutilities or costs for the users (i.e. users would be better off if the values of performance variables were reduced). The *average generalized transportation link cost*, or simply the *transportation link cost*, is a variable *synthesizing* (the average value of) the different performance variables *borne and perceived by the users* in travel related choice and, more in particular, in path choices (see section 4.3.4). Thus, the transportation link cost reflects the average users' disutility for carrying out the activity represented by the link. Other performance variables and costs, which cannot be associated to individual links but rather to the whole trip (path), will be introduced shortly.

Performance variables making up the transportation cost are usually non-homogeneous quantities. In order to reduce the cost to a single scalar quantity, the different components can be homogenized into a generalized cost applying reciprocal substitution coefficients β, whose value can be estimated by calibrating the path choice model (see section 4.3.4). For example, the generalized transportation cost, c_l, relative to the link l can be formulated as:

$$c_l = \beta_1 t_l + \beta_2 mc_l$$

where t_l is the travel time and mc_l is the monetary cost (e.g. the toll) connected with the crossing of the link. More generally, the link transportation cost can be expressed as a function of several link performance variables as:

$$c_l = \Sigma_n \beta_n r_{nl} \qquad (2.2.1)$$



LINK-PATH INCIDENCE MATRIX

O-D Pairs		1-4			2-4		3-4
Links \ Paths		1	2	3	4	5	6
1		1	1	0	0	0	0
2		0	0	1	0	0	0
3		1	0	0	1	0	0
4		0	1	0	0	1	0
5		1	0	1	1	0	1

Fig. 2.2.1 Example of a graph and link-path incidence matrix.

Different users may experiment and/or perceive transportation costs, which are different for the same link and the analyst does not know these costs. For example, the travel time of a certain road section is in general different for each vehicle that covers it, even under similar external conditions. Furthermore, two users experimenting the same travel time may have different perceptions of its disutility. Thus, the perceived link cost can be considered a random variable distributed among the users; the average value is the transportation link cost c_l.

There may be other "costs" both for the users (e.g. accident risks or tire consumption) and for the collectivity (e.g. noise and air pollution) associated to a link. It is usually assumed that these costs are not taken into account by the user in their travel-related choices and are not included in the perceived transportation cost. The transportation cost is, therefore, an *internal* cost, used for the simulation of the transportation system and, in particular, of travelers' choices. The other cost items represent an *external* cost, used for the design and the evaluation of projects. External costs are sometimes related to as impacts; they will be dealt with in section 2.2.5.

Different groups (or classes) of users may have different average transportation costs. This may be due to different performance variables (e.g. their speeds and travel times are different or they pay different fares) or to differences in the homogenization coefficients, β_n, (e.g. different time/money substitution rates corresponding to different incomes). In this case a link cost c^i_l can be associated with each user class i. In what follows, for simplicity of notation, the class index i will be taken as understood unless otherwise stated. Other considerations relative to users belonging to different classes will be made in Chapter 5.

Link performance variables and transportation costs can be arranged in vectors. The *performance vector*, $\boldsymbol{r}_l$, is made up by the n-th performance variable for each link, its components being r_{nl}. Analogously, the vector $\boldsymbol{c}$, whose generic component c_l is the generalized transport cost on the link l, is known as the *link cost vector*. Link performance and cost vectors have dimension $(n_L \times 1)$ where n_L is the number of links. The concepts of performance variables and generalized transportation cost can be extended from links to paths.

The *average performance variable of a path k*, z_{nk}, is the average value of that variable associated to a whole origin-destination trip, represented by a path in the graph. Some path performance variables are *link-wise additive*; i.e. their path value can be obtained as the sum of link values for all links making up the path.

Examples of additive path variables are travel times (the total travel time of a path is the sum of travel times over individual links) or some monetary costs, which can be associated to some or all individual links. An *additive path performance variable* can be expressed as the sum of link performance variables as:

$$z_{nk}^{ADD} = \sum_{l \in k} r_{nl} = \sum_{l} \delta_{lk} r_{nl} \tag{2.2.2a}$$

or in vector notation

$$\boldsymbol{z}_k^{ADD} = \Delta^T \boldsymbol{r}_n \tag{2.2.2b}$$

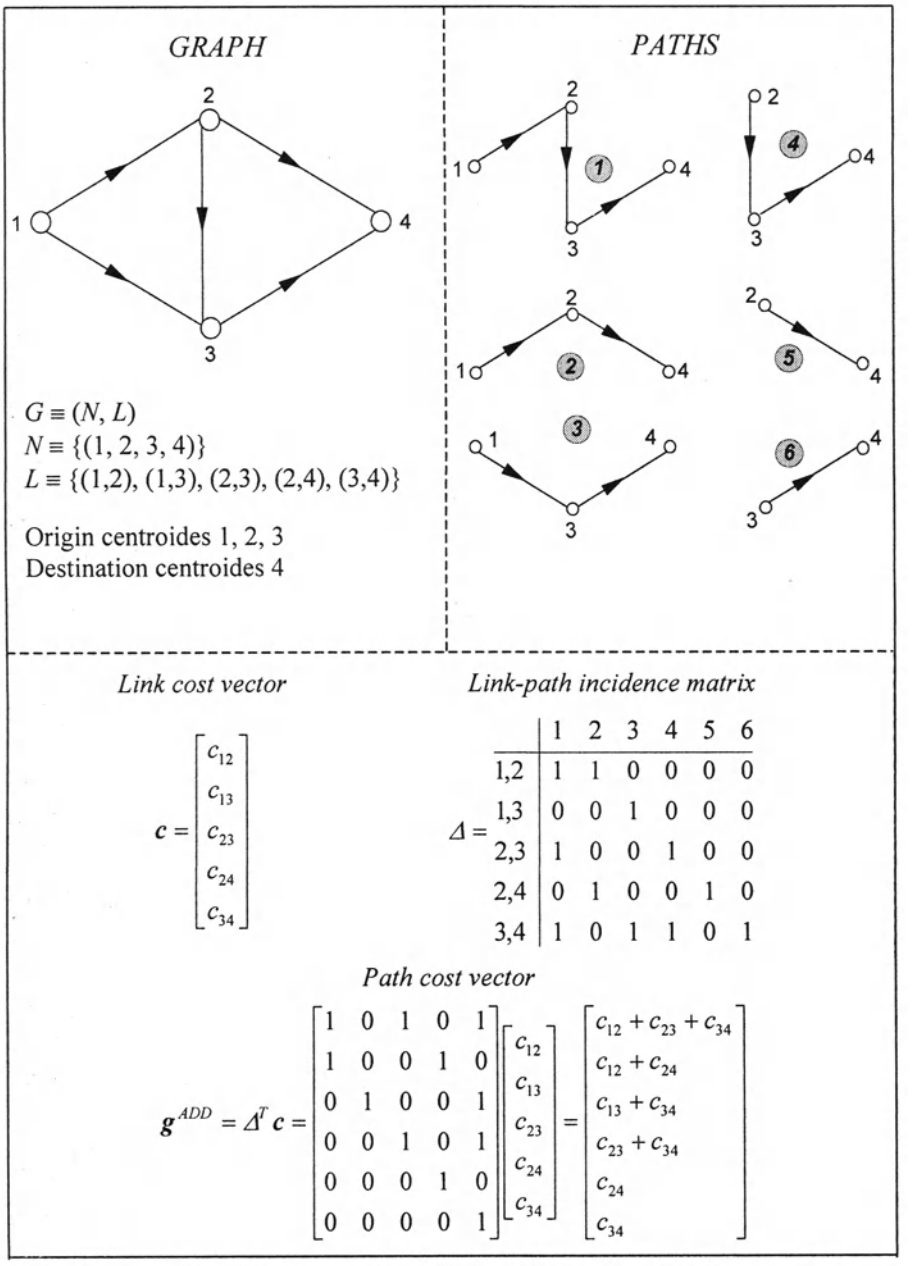


Fig. 2.2.2 Transportation network with link and path costs.

Other path performance variables are *non-additive*; i.e. they cannot be obtained as sum of link specific values. These variables are denoted by z_{nk}^{NA}. Examples of non-additive performance variables are the monetary cost in the case of tolls non-linearly proportional to the distance covered or the waiting time at stops for high-frequency transit systems, as will be seen below.

The *average generalized transportation cost* of a path k, g_k, is defined as a scalar quantity homogenizing in disutility units the different performance variables perceived by the users (of a given category) in making trip-related choices and, in particular, path choices.

The path cost in the most general case is made up of two parts: link-wise additive cost, g_k^{ADD}, and non-additive cost, g_k^{NA}, assuming that they are homogenous:

$$g_k = g_k^{ADD} + g_k^{NA}$$

The *additive path cost* is defined as the sum of the link-wise additive path performance variables:

$$g_k^{ADD} = \sum_n \beta_n z_{nk}^{ADD} \tag{2.2.3}$$

Under the assumption that the generalized cost depends linearly from performance variables, the additive path cost can be expressed as the sum of generalized link cost c_l.

The relationship between additive path cost and link costs can be expressed by combining eqns. (2.2.3), (2.2.2) and (2.2.1):

$$g_k^{ADD} = \sum_n \beta_n z_{nk}^{ADD} = \sum_n \beta_n \sum_l \delta_{lk} r_{nl} = \sum_l \delta_{lk} \sum_n \beta_n r_{nl} = \sum_l \delta_{lk} c_l \tag{2.2.4}$$

where, as stated, δ_{lk} is equal to one if the link l belongs to the path k, zero otherwise. The expression (2.2.4) can also be formulated in vector format by introducing the vector of additive path costs, $\boldsymbol{g}^{ADD}$, of dimensions ($n_P \times 1$):

$$\boldsymbol{g}^{ADD} = \boldsymbol{\Delta}^T \boldsymbol{c} \tag{2.2.5}$$

An example of the relationship (2.2.5) is depicted in Fig. 2.2.2.

The non-additive path cost, g_k^{NA}, includes non-additive path performance variables:

$$g_k^{NA} = \sum_n \beta_n z_{nk}^{NA}$$

Finally, the path cost vector, $\boldsymbol{g}$, of dimensions ($n_P \times 1$), can be expressed as:

$$\boldsymbol{g} = \boldsymbol{\Delta}^T \boldsymbol{c} + \boldsymbol{g}^{NA} \tag{2.2.6}$$

where $\boldsymbol{g}^{NA}$ is the non-additive path cost vector.

In many applications, the non-additive path cost vector is, or is assumed to be, null since this assumption simplifies the theoretical analysis and allows the use of efficient algorithms (e.g. implicit path enumeration) for the network assignment models, as discussed in Chapters 5 and 7 respectively.

2.2.3. Flows

A *link flow*, f_l, can be associated to each link l. Link flows, under the assumption of intra-period (within-day) stationarity, are the number of homogeneous units using the link (i.e. carrying out the trip phase represented by the link) in a time unit. Also in this case, the flow is properly a random variable whose average value is represented by the model. Several link flows can be associated to a given link depending on the homogeneous unit considered. *User flows* relate to users, such as travelers or goods possibly of different classes. *Vehicle flows* relate to the number of vehicles, possibly of different types such as automobiles, buses, trains, etc. If the link represents the crossing of a physical infrastructure (such as a road segment), the flow value, on the stationarity assumption, can be associated with each of its cross-sections (see Appendix 2A).

User flows are derived from demand models and influence supply performances; vehicle flows are usually associated with supply performance models. For individual modes, such as automobiles or trucks, user flows can be transformed quite straightforwardly into vehicle flows through average occupancy coefficients. For scheduled modes, such as trains, vehicle flows derive from the service schedule and are often treated as an input to the supply model. The link flow of the generic user class or vehicle type will be denoted by f_l^i.

Link performance variables and costs are often assumed to depend on *equivalent flows* associated with the links. In this case the flows of different user classes or vehicle types are homogenized to a reference class or type:

$$f_l = \sum_i w_i f_l^i \tag{2.2.7}$$

where w_i is the homogenization coefficient of the users of class i with respect to their influence on link performances. For example, for road flows, automobiles are usually the reference vehicle type ($w_i = 1$) and the other vehicle flows are transformed into equivalent auto flows with coefficients w_i. The latter ones are greater than one if the contribution to congestion of these vehicles is greater than that of the cars (buses, heavy vehicles, etc.), less then one in the opposite case (motorcycles, bicycles, etc.).

The *link flow vector*, $\boldsymbol{f}$, has dimensions ($n_L \times 1$), its generic component is the flow f_l on link l (see Fig. 2.2.3).

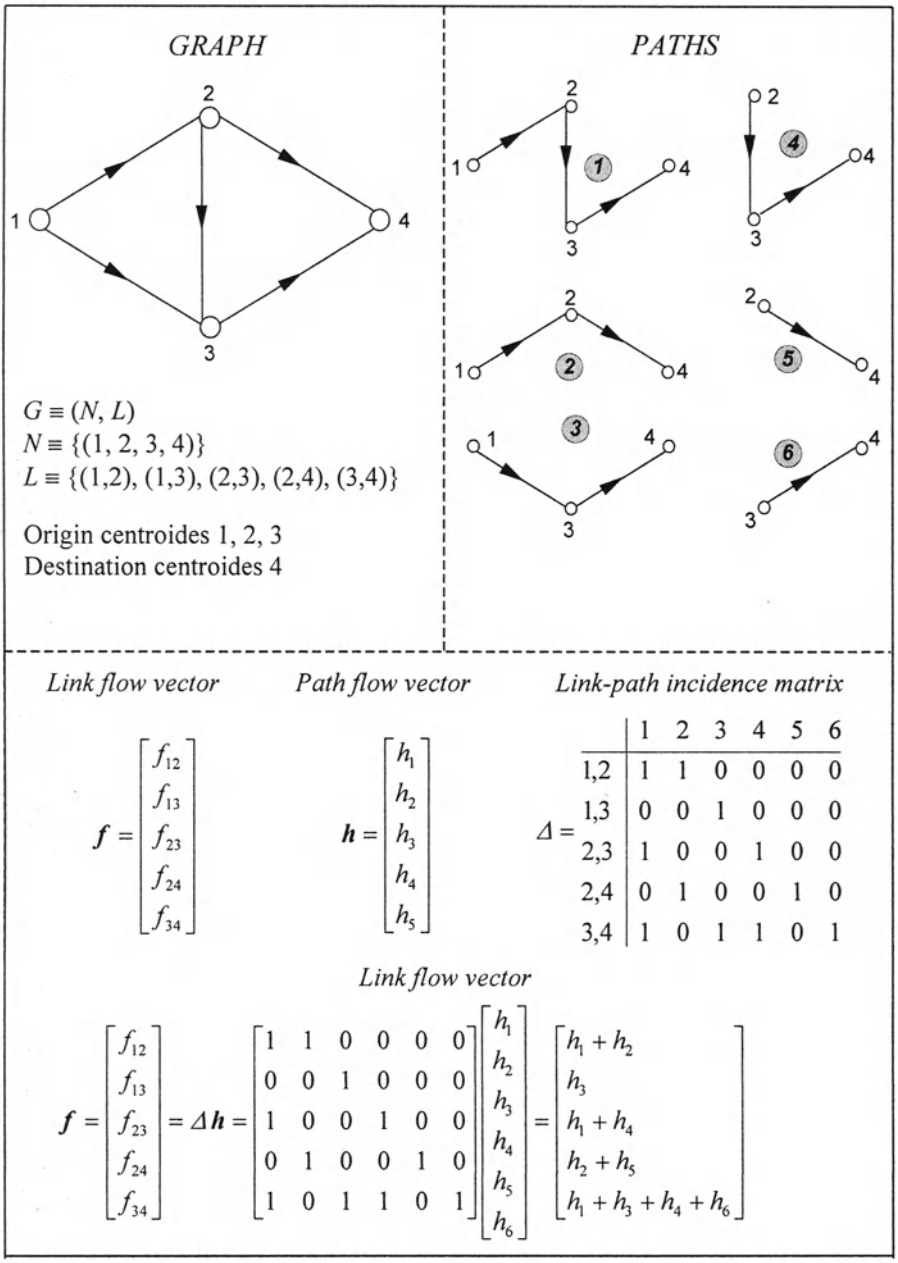


Fig. 2.2.3 Transportation network with link and path flows.

Flow variables can also be associated with paths. Under the within-day stationarity hypothesis, the average number of users, who in each sub-interval travel along each path, is constant. The average number of users, which in a time unit follow path k, is called the *path flow* h_k. If the users have different characteristics, i.e. belong to different classes, path flows per class i, h_k^i, can be introduced. Path flows of different user classes or vehicle types can be homogenized by means of coefficients w_i similar to those introduced for link flows; the equivalent path flow is obtained as:

$$h_k = \Sigma_i w_i h_k^i \tag{2.2.8}$$

Using a vector format, can be defined the *path flow vectors* $\boldsymbol{h}^i$ and $\boldsymbol{h}$; they are column vectors of dimensions ($n_P \times 1$) (see Fig. 2.2.3).

The user flow following each path can also be seen as a random variable since, in general, it can vary over different observation periods of the system.

There is clearly a relationship between link and path flows. The flow on each link l, in fact, can be obtained as the sum of the flows on the various paths containing that link. This relationship can be expressed by using the elements δ_{lk} of the link-path incidence matrix as:

$$f_l = \Sigma_k \delta_{lk} h_k \tag{2.2.9}$$

or in matrix terms:

$$\boldsymbol{f} = \Delta \boldsymbol{h} \tag{2.2.10}$$

Equation (2.2.9), or (2.2.10), expresses the way in which path flows induce flows on individual links. For this reason it will be referred to as the (*static*) *Network Flow Propagation* (*NFP*) or *Network Loading* (*NL*) model (see Fig. 2.1.1). Note that the linear structure of equation (2.2.9) depends crucially on the assumption of intra-period stationarity (within-day static model); if this assumption is removed, the *NL* model becomes significantly more complicated as will be seen in Chapter 6.

Furthermore, note the difference between the relationships connecting link and path costs and flows. As far as the costs are concerned, the additive path cost is given by the sum of the component link costs as expressed in the equation (2.2.3) or (2.2.4). On the other hand, the link flow is obtained by the sum of flows on the paths, which include that link (equation (2.2.9) or (2.2.10)).

2.2.4. Link performance and cost functions

Link performance attributes generally depend on the physical and functional characteristics of the facility and/or the service involved in the trip phase represented by the link itself. Typical examples are the travel time on a road section depending on its length, alignment, allowed speed or the waiting time at a bus stop depending on the headway between successive bus arrivals. When several travelers or vehicles use the same facility, they may interact with each other influencing link

performances. This phenomenon is known as *congestion.* Typically, the effects of congestion on link performances increase as the flow increases. For instance, the larger the flow of vehicles traveling along a road section, the more likely faster vehicles will be slowed by slower ones, thus increasing the average travel time. Moreover, the larger the flow arriving at an intersection, the larger the average waiting time; the larger the number of users on the same train, the lower the riding comfort.

In general, congestion effects are such that the performance attributes of a given link may be influenced by the flow on the link itself and by flows on other links.

Link performance functions relate the generic link performance attribute, r_{nl}, to physical and functional characteristics of the link, arranged in a vector $\boldsymbol{b}_{nl}$, and to the equivalent flow on the same link and, possibly, on other links, arranged in the vector $\boldsymbol{f}$:

$$r_{nl} = r_{nl}(\boldsymbol{f}; \boldsymbol{b}_{nl}, \gamma_{nl}) \tag{2.2.11}$$

where γ_{nl} is a vector of parameters used in the function.

Since the generalized transportation cost of a link, c_l, is a linear combination of link performance attributes, as expressed by equation (2.2.1), *link cost functions*[1] can be expressed as functions of the same parameters:

$$c_l = c_l(\boldsymbol{f}; \boldsymbol{b}_l, \gamma_l) \tag{2.2.12}$$

Given the relevance of congestion effects on the analysis of transportation systems, in the following link performance and cost functions will explicitly express their dependence of link flows as $r_{nl}(\boldsymbol{f})$ and $c_l(\boldsymbol{f})$ respectively; vectors $\boldsymbol{b}_l$ and γ_l will be understood.

Link performance and cost functions may have some mathematical properties, which will be used in Chapter 5 to study the properties of supply-demand interaction models and in Chapter 7 to analyze the convergence of their solution algorithms. It is sometimes useful to separate the link cost, and therefore the functions simulating the component performance attributes, in two parts. The variable cost $cv_l(\boldsymbol{f})$ includes those performance attributes, usually travel and/or waiting time, which vary significantly because of the congestion effect and are regarded as functions of equivalent flows. The fixed cost c_{0l} includes those performance attributes, e.g. tolls, which are independent of link flows.

In general, therefore, the link cost function can be expressed as:

$$c_l(\boldsymbol{f}) = c_{0l} + cv_l(\boldsymbol{f}) \tag{2.2.13}$$

assuming that c_{0l} and cv_l are expressed in homogeneous (disutility) units.

Performance and cost functions can be classified as *separable* and *non-separable* across link. In the former case, the performances and cost variables of a link depend exclusively on the (equivalent) flow on the link itself:

$$c_l(\boldsymbol{f}) = c_l(f_l)$$

In the latter case, they also depend on the flow on other links. Examples of both types of function will be given in the following sections.

The *cost function vector*, $\boldsymbol{c}(\boldsymbol{f})$, of dimensions $(n_L \times 1)$, is obtained by ordering the n_L functions of the individual network links:

$$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f}) \tag{2.2.14}$$

The Jacobian matrix $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$ of the functions vector, $\boldsymbol{c}(\boldsymbol{f})$, has dimensions $(n_L \times n_L)$ and can be expressed as:

$$\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})] = \begin{vmatrix} \dfrac{\partial c_1}{\partial f_1} & \cdots & \dfrac{\partial c_1}{\partial f_{n_L}} \\ \cdots & \dfrac{\partial c_i}{\partial f_i} & \cdots \\ \dfrac{\partial c_{n_L}}{\partial f_1} & \cdots & \dfrac{\partial c_{n_L}}{\partial f_{n_L}} \end{vmatrix} \tag{2.2.15}$$

Separable and non-separable cost functions can be characterized with reference to their Jacobians. In the separable case, the Jacobian will be a diagonal matrix: $\partial c_i/\partial f_j = 0$, $\forall\ i \neq j$. Symmetric cost functions have a symmetric Jacobian matrix: $\partial c_i/\partial f_j = \partial c_j/\partial f_i$, $\forall\ i, j$; i.e. the cost variation on link l, due to a flow variation on link j, is equal to the cost variation on link j, due to a flow variation on link i. Asymmetric cost functions, instead, have an asymmetric Jacobian. Separable cost functions are clearly a special case of symmetric functions. A further special case is the cost function vector of an uncongested network. In this case the cost functions are independent of the flows, so the partial derivatives of (2.2.15) are all equal to zero and the Jacobian is null.

2.2.5. Impacts and impact functions

Design and evaluation of transportation systems, in addition to performance variables perceived by the users, require the modeling of impacts borne by the users, but not perceived in their mobility choices, and of impacts on non-users. Examples of the first type include indirect vehicle costs (e.g. tire or lubricant, vehicle depreciation, etc.) and accident risks with their consequences (deaths, injuries, material damages). The impacts for non-users include those for other subjects directly involved in the transportation system, such as costs and revenues for the producers of transportation services, and impacts "external" to the transportation system (or market). Examples of externalities are the impacts on the real estate market, urban structure, or on the environment such as noise and air pollution. The

mathematical functions relating these impacts to physical and functional parameters of the specific transportation systems and, in some cases, to link flows are called *impact functions*. Often these functions are named with respect to the specific impact they simulate (e.g. fuel consumption functions or pollutants emission functions). Some impacts can be associated with individual network links and depend on the flows, $e_l(f)$. Link-based impact functions are usually included in transportation supply models; see Fig. 2.1.1. Some impact functions may be quite elementary while others may require complex systems of mathematical models. Examples of link-based impact functions are those related to air and noise pollution due to vehicular traffic. Some impact functions will be discussed in Chapter 10 in the context of evaluation of transportation system projects.

2.2.6. General formulation

A *transportation network* consists of the set of nodes N, the set of links L, the vector of link costs $\boldsymbol{c}$, which depend on the vector $\boldsymbol{r}$ of link performances, the vector $\boldsymbol{g}^{NA}$ of non-additive path costs and the vector $\boldsymbol{e}$ of relevant impact variables: (N, L, $\boldsymbol{c}$, $\boldsymbol{g}^{NA}$, $\boldsymbol{e}$). For the sake of simplicity, the set of relevant paths, as will be defined in Chapter 6, is not indicated. For congested networks, the link cost vector is substituted by the flow-dependent cost functions $\boldsymbol{c}(\boldsymbol{f})$; the same holds for flow-dependent internal and external impacts, $\boldsymbol{e}(\boldsymbol{f})$, while the non-additive costs vector, $\boldsymbol{g}^{NA}$, is usually assumed to be independent of the flows. In this case the abstract transportation network model can be expressed as (N, L, $\boldsymbol{c}(\boldsymbol{f})$, $\boldsymbol{g}^{NA}$, $\boldsymbol{e}(\boldsymbol{f})$). Performance variables and functions are not explicitly mentioned, as they are included in the generalized transportation cost functions.

The set of relationships connecting path costs to path flows is known as the *supply model*. The supply model can therefore be formally expressed combining the equations (2.2.6), (2.2.10) and (2.2.14) into a relationship connecting path flows to path costs:

$$\boldsymbol{g}(\boldsymbol{h}) = \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}) + \boldsymbol{g}^{NA} \tag{2.2.16}$$

where it is assumed that non-additive path costs, if any, are not affected by congestion. Link characteristics can be obtained through performance, cost and impact functions for the link flows corresponding to the path flow vector. Clearly the model (2.2.16) expresses the abstract congested network model described in the previous sections. The same type of models can be used to describe other systems such as electrical or hydraulic networks.

2.3. Applications of Transportation Supply models

Network models and related algorithms are powerful tools for modeling transportation systems. A network model is a simplified mathematical description of the physical phenomena relevant for the analysis, design and evaluation of a given system. Thus transportation network models depend on the purpose for which they are used.

Building a network model usually requires a sequence of operations whose general criteria will be described in the following. A schematic representation of the main activities in the case of a bi-modal supply system (road and transit urban systems) is depicted in Fig. 2.3.1.

In the most general case, a supply network model is built through the following phases:

a) delimitation of the study area;
b) zoning;
c) selection of relevant supply elements (basic network);
d) graph construction;
e) identification of performance and cost functions;
f) identification of impact functions.

Phases a), b) and c) relate to the relevant supply system definition. They are described respectively in section 1.2.1 of Chapter 1 and will not be repeated here. The rest of this section will introduce some general considerations related to phases d), e) and f) for a generic system. Specific models will be described separately for two different types of transportation systems: continuous services (such as road), in section 2.3.1, and scheduled services (such as train or buses), in section 2.3.2.

The construction of a *transportation graph* requires the definition of the relevant trip phases and events (links and nodes) that depend on the physical system to be represented. Important nodes in transportation graphs are the so-called *centroid nodes*. They correspond to the events of beginning and ending a trip in a given zone. As was seen in section 1.2.1, the centroids can approximate the internal points within a traffic zone. In general the zone centroid is a *fictitious node*, i.e. a node which does not correspond to any specific location but which represents the set of the points of the zone where a trip can start or end. For this reason, a zone centroid is placed "barycentrically" with respect to such points or to some proxy variables (e.g. the number of households or workplaces). In principle, different centroid nodes may be associated to different trip types (e.g. origin and destination centroids). In other cases, centroids represent the places of entry into or exit from the study area for the trips, which are partly carried out within the system (*cordon centroids*). In this case they are usually associated with physical locations (road sections, airports, railway stations, etc.).

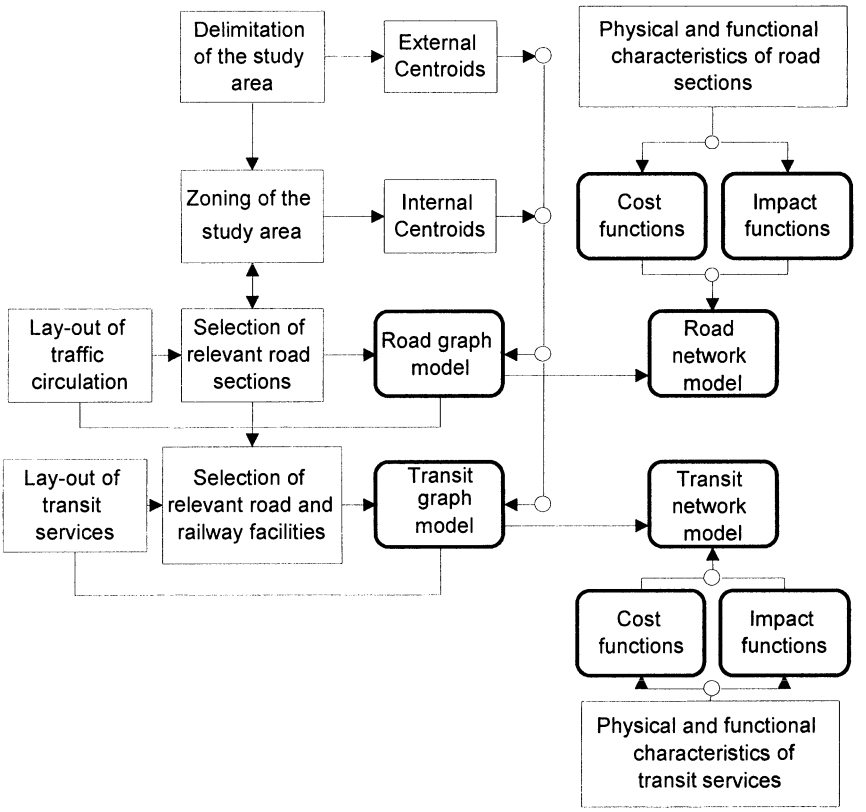


Fig. 2.3.1 Functional phases for the construction of an urban bi-modal network model.

A graph usually includes *links* of different types: *real links* and *connectors*. Real links represent trip phases corresponding to "physical" components (infrastructures or services), such as traversing a road section or riding a train between two successive stations. When centroid nodes do not correspond to a physical element, *connector links* are introduced into the graph. These links represent the trip phase between the terminal point (zone centroid) and a physical element of the network. In the remainder of this section, links will be referred to according to the phase (activity) of the trip or to the infrastructure or service, which allows that activity. For example, there will be road links, transit line links, and waiting links at stops.

A transportation graph will have different levels of complexity, depending on the system being represented and the details needed for its representation. In general, it can be said that short-term, or operational, projects, such as a road circulation plan or the design of transit lines, require a very detailed representation of the real system. On the other hand strategic, or long-term, projects usually require less

detailed, larger-scale graphs both because of the geographical size of the area and of the number of elements included in the system.

As will be seen soon, different graphs can be associated with the same basic network, depending on the aim of the model.

Graphs can also represent transportation infrastructures; in general, infrastructure graphs are not used directly for system models, but rather they are referred to during the construction of service graphs. User flows and supply performances depend on the transportation services using the infrastructures rather than on the infrastructures themselves.

To specify *link performance* and *cost functions* for a transportation network, the operational characteristics of its elements are needed. Performance functions used in practice sometimes derive from an explicit mathematical model, such as *queuing models*, for so-called *barrier systems* (motorway toll-barriers, road intersections, air and maritime terminals, etc.), or *traffic flow theory models*, for continuous systems. When the formulation of an explicit mathematical model, even in a simplified form, turns out to be particularly complex, "descriptive" models are used instead. These models are statistical relationships between performance attributes and explanatory variables. Examples of both types of performance functions will be given in the following two sections and in Appendix 2.A.

Both explicit and descriptive models include unknown parameters, vectors γ_n and γ in expressions (2.2.11) and (2.2.12) respectively, that should be calibrated for the specific supply system. To estimate the parameters of theoretical models or to specify the functional form and estimate the parameters of descriptive models, traditional estimation methods can be used. In particular, least square estimators are often adopted to this purpose. In many applications cost functions calibrated in similar contexts are transferred to the system under study to save on times and costs.

2.3.1. Supply models for continuous service systems

Continuous and simultaneous services are available at every instant and can be accessed from a very large number of points. Typical examples are individual modes such as cars and pedestrians using road systems.

2.3.1.1. Graph models

In graphs representing road systems, nodes are usually located at the intersections between road segments included in the supply model. Nodes can also be located where significant variations of the geometrical and/or functional characteristics of a single segment occur (such as changes in a road cross-section and lateral friction). Intersections with secondary roads not included in the "base network", however, are not represented by nodes. Links usually correspond to connections between nodes allowed by the circulation scheme. Therefore, a two-way road will be represented by two links going in opposite directions, while a one-way road will have a single link going in the allowed direction. Fig. 2.3.2 shows the graph representing the part of urban road network shown in Fig. 1.2.2.

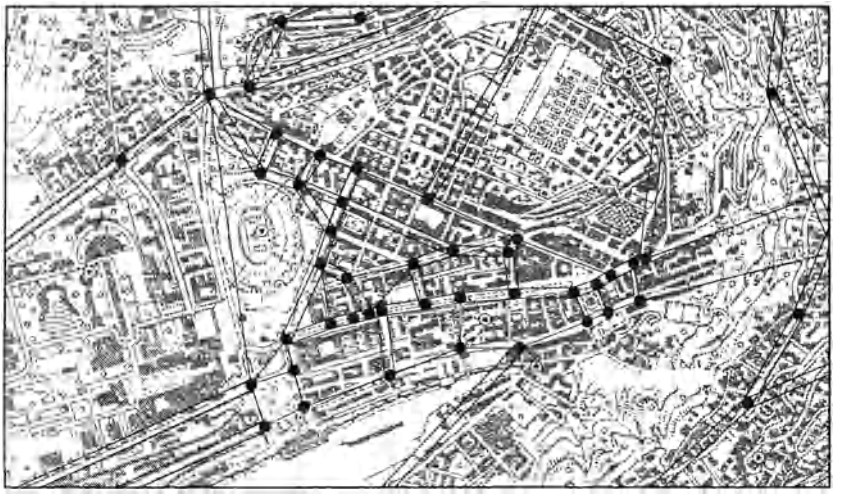


Fig. 2.3.2 Example of a graph representing part of an urban road system.

The level of detail of the road system depends on the purpose of the model. This is especially true for road intersections. In a coarse representation, a road intersection is usually represented by a single node where the access links converge. Alternatively, it is possible to adopt a more detailed representation that distinguishes different turning movement and excludes non-permitted turns (if any). Such a representation can be obtained by using a larger number of nodes and links. Fig. 2.3.3 shows the two possible representations of a four-arm road intersection. Note that in the single node representation, paths requiring the left turn (4-5-2) cannot be excluded if this turning movement is not allowed; furthermore, different waiting times cannot be assigned to maneuvers with different green phase durations, such as the right turn (4-5-3). Both of these possibilities are allowed by the detailed representation.

Parking is another element of a road system that can be represented with different levels of detail. In detailed road graph, trip phases corresponding to parking can be represented with different links for different parking facilities available in a given zone, see Fig. 2.3.4. *Parking links* can be connected through pedestrian links to the centroid of the zone where they are located in, and to the centroids of traffic zones within walking distance. In less detailed graphs, parking is included in connector links; in this case however it is not possible to simulate congestion and different parking policies.

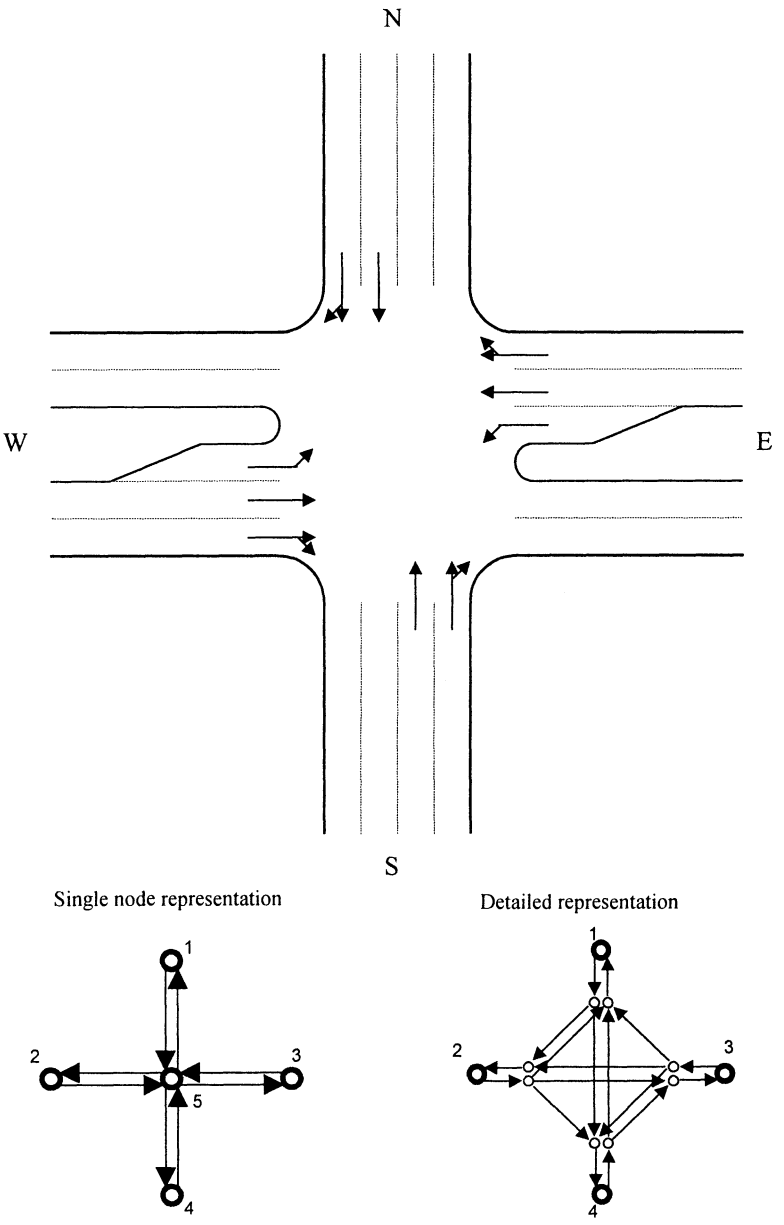


Fig. 2.3.3 Graphs for a road intersection.

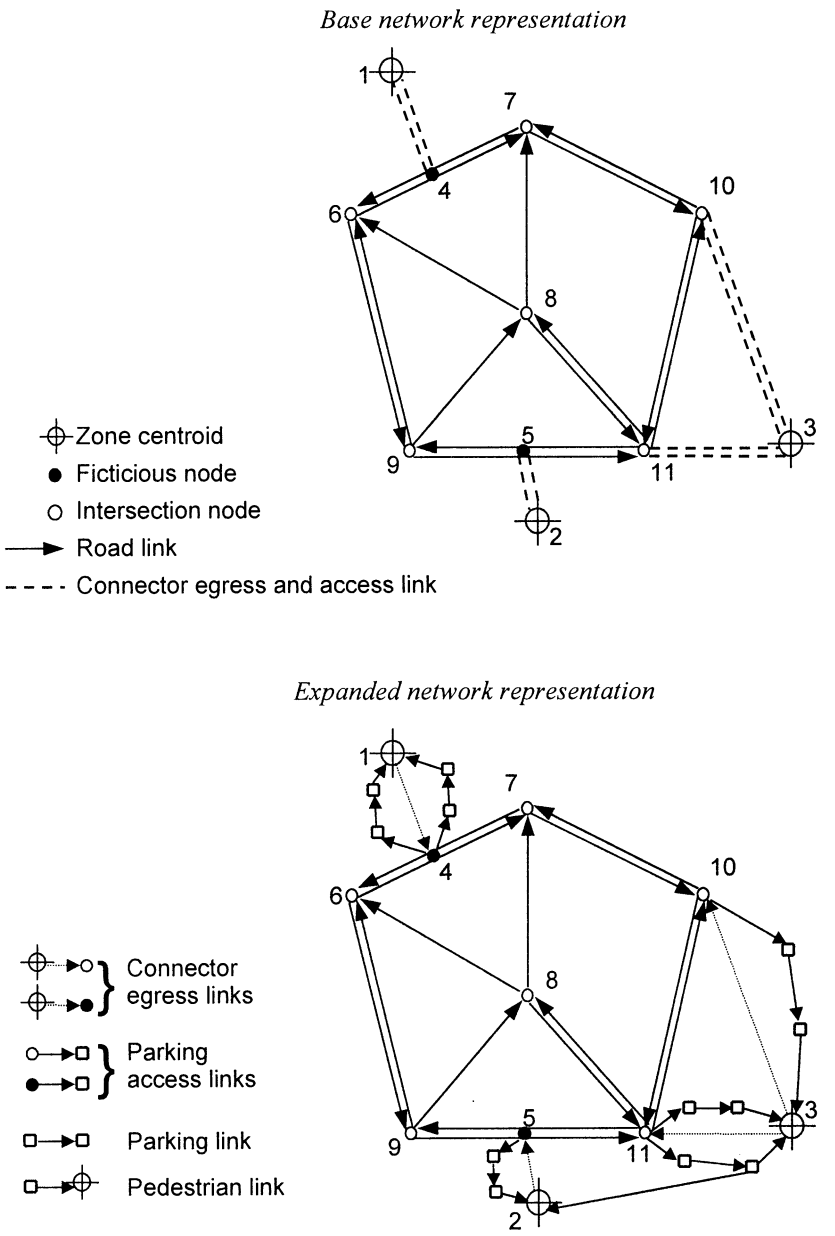


Fig. 2.3.4 Explicit representation of parking supply.

2.3.1.2. Performance and cost functions

The *generalized transportation cost* of a road link is usually made up by several performance attributes. For example, three attributes can be selected: travel time along the section, waiting time (e.g. at the final intersection, at the tollbooth, etc.) and monetary cost. In this case, the cost function can be obtained as the sum of three performance functions:

$$c_l(\boldsymbol{f}) = \beta_1 tr_l(\boldsymbol{f}) + \beta_2 tw_l(\boldsymbol{f}) + \beta_3 mc_l(\boldsymbol{f}) \tag{2.3.1}$$

where:

$tr_l(\boldsymbol{f})$ is the function relating the running time on link l to the flow vector;
$tw_l(\boldsymbol{f})$ is the function relating the waiting time on link l to the flow vector;
$mc_l(\boldsymbol{f})$ is the function relating the monetary cost on link l to the flow vector.

The dependence on physical and functional variables, $\boldsymbol{b}_l$, and parameters, γ, has been omitted for simplicity sake. Note that in equation (2.3.1) it has been assumed that homogenization coefficients may be different for the different time components. Furthermore, not all of the components in (2.3.1) are present for each link; for example, if the link represents only the waiting time for a maneuver, tr_l and mc_l are zero, and the same consideration is true for monetary costs and waiting times on most pedestrian links.

In the most general case, the monetary cost term mc_l includes the cost items that are perceived by the user. Since the users do usually not perceive other consumptions (mineral oil, tires, etc.), in applications monetary costs are usually identified as the toll (if any) and the fuel consumption:

$$mc_l = mc_{toll} + mc_{fuel}(\boldsymbol{f})$$

The latter depends on the specific consumption (liters/km), which can vary in relation to the average speed and, therefore, to the congestion level. In practice, these variations are sometimes ignored and the monetary cost is calculated as a function of the toll and the average unitary consumption.

Listing all the performance functions that can be adopted for the elements of different continuous service systems is beyond the scope of this book. In the following, some examples of performance functions for typical links of road networks will be given. It should also be stressed that, consistently with the assumption of intra-period stationarity, stationary traffic flow variables and results will be used (see Appendix 2.A.). Other delay functions are described in Appendix 2.A. in the context of traffic flow and queuing theories.

a) *Motorway links*

On motorway links flow conditions are typically uninterrupted and it is assumed that the waiting time component is negligible since it occurs on those sections (ramps, tollbooths, etc.), which are usually represented by different links.

Link travel time is usually obtained through empirical statistical relationships. One of the most popular expressions, referred to as the BPR cost function, has the following specification:

$$tr_l(f_l) = \frac{L_l}{v_{0l}} + \gamma_1 \left(\frac{L_l}{v_{cl}} - \frac{L_l}{v_{0l}} \right) \left(\frac{f_l}{Q_l} \right)^{\gamma_2} \tag{2.3.2}$$

where:

L_l is the length of link l;
v_{0l} is the free-flow average speed;
v_{cl} is the average speed with flow equal to capacity;
Q_l is the link capacity, i.e. the average maximum number of equivalent vehicles that can travel along the road section in a time unit. Capacity is usually obtained as the product of the number of lanes on the link l, N_l, and lane capacity, Q_{ul};
γ_1, γ_2 are parameters of the function.

From eqn (2.3.2) it can be noted that, in the case of motorways, cost functions are separable. The influence of flows on the performances of other links (e.g. the opposite direction or entrance/exit ramps) is significantly reduced by the characteristics of the infrastructure (divided carriageways, grade-separated intersections, etc.).

The values of v_{0l}, v_{cl} and Q_l depend on the geometric and functional characteristics of the section (width of lanes, shoulders and median strips, bend radiuses, longitudinal slopes, etc.). Typical values can be found in different sources; the Highway Capacity Manual (HCM) is the most complete and systematic (see bibliographic note). The parameters γ_1 and γ_2 are typically estimated on empirical data.

Fig. 2.3.5 shows a diagram of eqn (2.3.2) for different parameter values. Note that this function associates a travel time to the link also when flows are above the link capacity (over-saturation), even though such flows are not possible in reality. However, in applications over-saturation is often allowed for reasons connected with mathematical properties and solution algorithms of static equilibrium assignment models (see Chapters 5 and 7). From a computational point of view, the over-saturation assumption should not impact significantly the results if the value of parameter γ_2, i.e. the delay penalty due to capacity overloading, is large enough.

Values of γ_2 are typically much larger than one; i.e. the function is more-than-linear in flow/capacity ratios. This phenomenon is rather frequent in congested

systems. Furthermore, as will be seen in Chapter 5, at equilibrium, users are distributed among paths in order to have equal (perceived) costs and, therefore, it is unlikely that the resulting flows are significantly higher than capacities. It should also be noted that, as shown in Appendix 2.A., if the flow is close to the capacity, resulting instability challenges the within-day stationarity assumptions and the cost functions adopted. In this sense, delay functions should be considered as "penalty" functions preventing major over-saturation, rather than estimates of actual travel times.

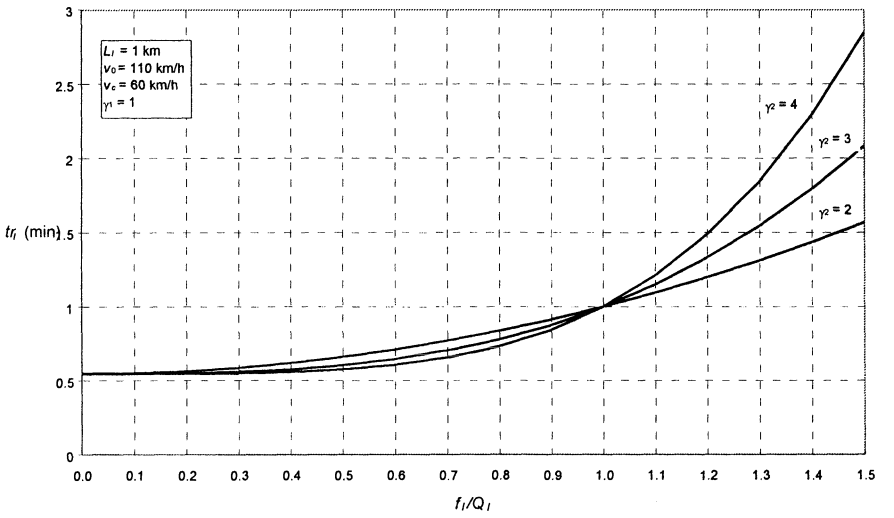


Fig. 2.3.5 Motorway travel time function (2.3.2) for different values of some parameters.

b) *Extra-urban road links*

Users travelling on an extra-urban road behave differently according to the number of lanes available for each direction: single lane (two-lane arterial) or two or more lanes (four-lane arterial, six-lane arterial, etc.).

In the former case, the capacity and travel conditions in each direction are not influenced by the flow in the opposite direction. For this type of road, the same formula (2.3.2) described for motorway links can be used, although with different parameters. These can again be deduced from capacity manuals, such as the HCM, or from other specific empirical studies.

In the case of roads with one lane in each direction, link performances depend on the flow in both directions since the overtaking maneuvers are not always possible and, consequently, the vehicles can reduce the average speed. In practice, it is often assumed that the link capacity has a value common to both the directions and the travel time function is modified as follows:

$$tr_l(f_l, f_{l^*}) = \frac{L_l}{v_{0l}} + \gamma_1 \left(\frac{L_l}{v_{cl}} - \frac{L_l}{v_{0l}} \right) \left(\frac{f_l + f_{l^*}}{Q_{ll^*}} \right)^{\gamma_2} \tag{2.3.3}$$

where, apart from the symbols introduced previously, the link in the opposite direction is denoted by l^* and the overall capacity in both directions by Q_{ll^*}.

c) *Toll-barrier links*

In the case of links representing queuing systems, it is assumed that the average waiting time is the only significant time performance variable. In simple cases (e.g. a link corresponds to all tolling lanes), the average under-saturation waiting time can be obtained by using a stochastic queuing model:

$$tw_l^u(f_l) = T_s + (T_s^2 + \sigma_s^2) \cdot \frac{f_l}{2} \cdot \frac{1}{1 - f_l/Q_l} \tag{2.3.4}$$

where:

T_s is the average service time for each toll-lane;
σ^2_s is the variance of the service time at the pay-point;
$Q_l = N_l/T_s$ is the link (toll-barrier) capacity equal to the product of the number of lanes (N_l) by the capacity of each lane ($1/T_s$).

Expression (2.3.4) is derived from the assumption of a queuing system *M/G/1* (∞, *FIFO*) with Poisson arrivals and general service time (see section 2.A.2.3).

The values of T_s and σ^2_s depend on various factors such as the tolling structure (fixed, variable) and the payment method (manual, automatic, etc.)

Note that the average waiting time obtained through equation (2.3.4) is larger than the average service time T_s even though the arriving flow is lower than the system's capacity. This effect derives from the presence of random fluctuations in the headways between user arrivals and the service times. For this reason, the delay expressed by (2.3.4) is known as "stochastic delay".

Note also that the average delay computed with equation (2.3.4) tends to infinity as the flow f_l tends to the capacity (i.e. if f_l/Q_l tends to one). This would be the case if the arrivals flow f_l remained equal to the capacity for an infinite time, which does not occur in reality. In order to avoid unrealistic waiting times and for reasons of theoretical and computational convenience, two different methods can be adopted. The first, and less precise, method assumes that equation (2.3.4) holds for flow values up to a fraction α of the capacity, for example $f_l \leq 0.95\ Q_l$. For higher values, the curve is extended following its *linear approximation*, i.e. in a straight line passing through the point of coordinates $\alpha\ Q_l$, $tw(\alpha\ Q_l)$ with angular coefficient equal to the derivative of (2.3.4) computed at this point:

$$tw_l(f_l) = tw_l(\alpha Q_l) + K(f_l - \alpha Q_l) \quad \text{for } f_l \geq \alpha Q_l \tag{2.3.5}$$

with

$$K = \frac{T_s^2 + \sigma_s^2}{2} \cdot \frac{1}{(1-\alpha)^2}$$

Fig. 2.3.6 shows the relationships (2.3.4) and (2.3.5) for some values of the parameters.

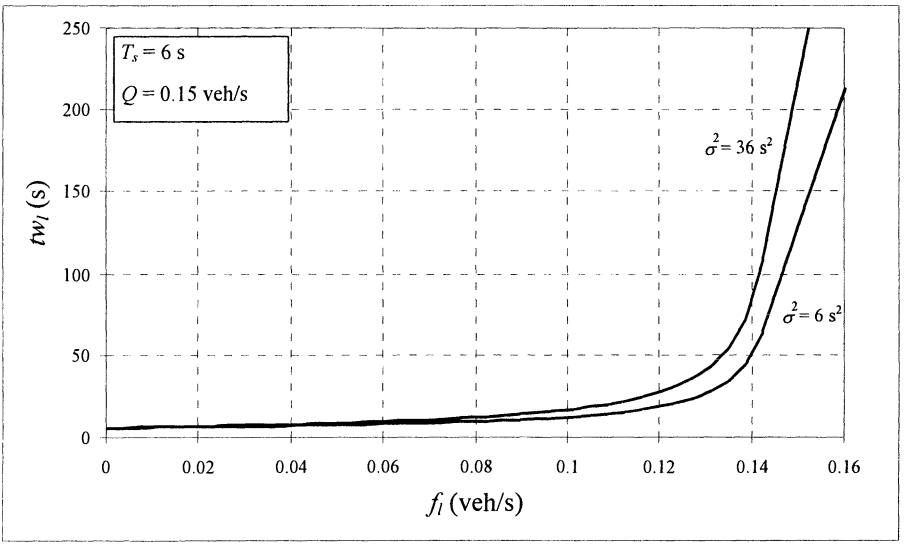


Fig. 2.3.6 Waiting time functions (2.3.4) and (2.3.5) at toll-barrier links.

A more rigorous method is based on the calculation of over-saturation delay using a deterministic queuing model with arrival rate equal to f_l, deterministic service times equal to T_s and over saturation period equal to the reference period duration T (see section 2.A.2.2). The deterministic average (over-saturation) delay tw^d_l is then equal to:

$$tw^d_l = T_s + \left(\frac{f_l}{Q_l} - 1\right)\frac{T}{2} \tag{2.3.6}$$

which, for a *given value* of capacity, is a linear function of the arrivals flow f_l.

Note that in this case the assumption of intra-period stationarity is challenged since even if the arrival flow rate, f_l, and the capacity, $1/T_s$, are constant over the whole reference period T, the waiting time is different for users arriving in different instants of the reference period. In static models it is assumed that users perceive

the average waiting time. Intra-period dynamic models, discussed in Chapter 6, remove this assumption.

The average delay, tw_l, can be calculated by combining the stochastic under-saturation average delay, tw^u_l, expressed by (2.3.4) with the deterministic average over-saturation delay, tw^d_l, expressed by (2.3.6). The combined delay function is such that the deterministic delay function is its oblique asymptote; see Fig. 2.3.7. The following equation results:

$$tw_l(f_l) = T_s + (T_s^2 + \sigma^2)\frac{f_l}{2} + \frac{T}{4}\left\{\frac{f_l}{Q_l} - 1 + \left[\left(\frac{f_l}{Q} - 1\right)^2 + \frac{4(f_l/Q_l)}{Q_l T}\right]^{1/2}\right\} \quad (2.3.7)$$

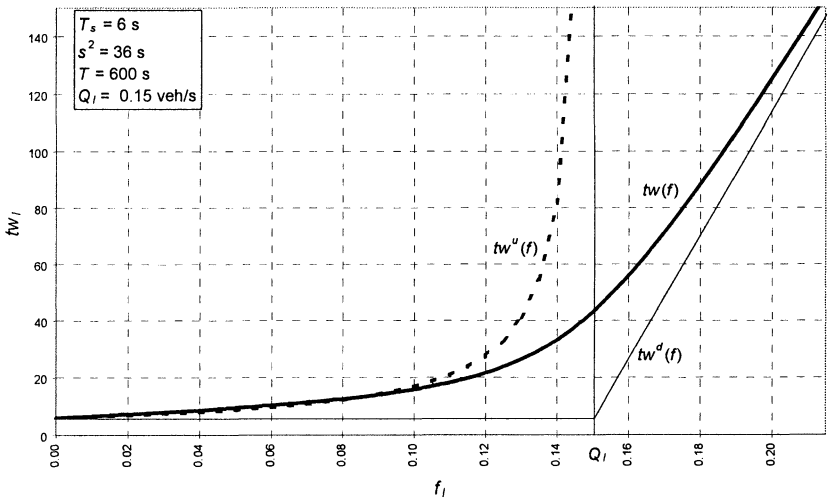


Fig. 2.3.7 Under- and over-saturation waiting time functions for toll barrier links.

d) *Urban road links*

Links representing urban road sections are often rather short (few hundred meters) and the average travel speed is not significantly influenced by the flow, both because of the short distance between two successive intersections, and because of the low speed limits.

The *running time* tr_l can be obtained through "descriptive" statistical models. Some models express the average speed on the link as a function of physical and functional parameters and of traffic flows:

$$tr_l(f_l) = \frac{L_l}{v_l(f_l)} \quad (2.3.8)$$

An example of estimation of $v_l(f_l)$ is the following empirical expression calibrated in some Italian urban areas (see bibligraphic note):

$$v_l = 31.1 + 2.8\, NW_l - 1.2\, S_l - 12.8\, WND_l^2 - 10.4\, CH_l - 1.4\, INT_l + \\ -[0.000053 + 0.000123\, X_l]\, (f_l / NW_l)^2 \tag{2.3.9}$$

where:

NW_l is the "net" width, i.e. the road width in each direction, reduced by the space occupied by parked vehicles (in meters);
S_l is the average slope in percentage units (%);
WND_l is the level of windingness in a scale [0, 1];
CH_l is the level of circulation hindrance due to pedestrians and parking movements in a scale [0, 1];
INT_l is the number of secondary intersections per kilometers;
X_l is a dummy variable equal to 1 if the road does not allow overtaking, zero otherwise;
f_l is the link flow in vehicles/h.

If a single link represents both the running along a road segment and the waiting at the final intersection its cost function will include both the components tr_l and tw_l, the latter discussed in the following subsection.

e) *Intersection links*
The *average waiting time*, tw_l, at an intersection can be computed by using theoretical and/or empirical formulae obtained for different types of regulation.

The simplest case is that of a *signal controlled intersection* not interacting with adjacent ones (*isolated intersection*), without lanes reserved for right or left turns. In this case, the average waiting time depends on the ratio between the flow f_l on the approach corresponding to the link l and the capacity of the approach itself Q_l. The latter is the average value of the maximum number of vehicles that can cross the approach in a time unit.

The capacity of the approach can be expressed as a fraction μ of the saturation flow, S:

$$Q = \mu S \qquad \mu = G/T_c$$

where:

$\mu = G/T_c$ is the *effective green ratio* i.e. the ratio between the effective green duration G (green + yellow – lost time) and the duration T_c of the *traffic-light cycle* (green + yellow + red);

S is the *saturation flow* of the approach; i.e. the maximum number of equivalent vehicles that in a time unit could cross the stop line under continuous green light ($\mu = 1$). Alternatively, the saturation flow can be defined as the maximum discharge rate across the stop-line that can be sustained by a continuous queue during the green-yellow time.

Fig. 2.3.8 gives a graphical illustration of the quantities associated with a traffic-light cycle.

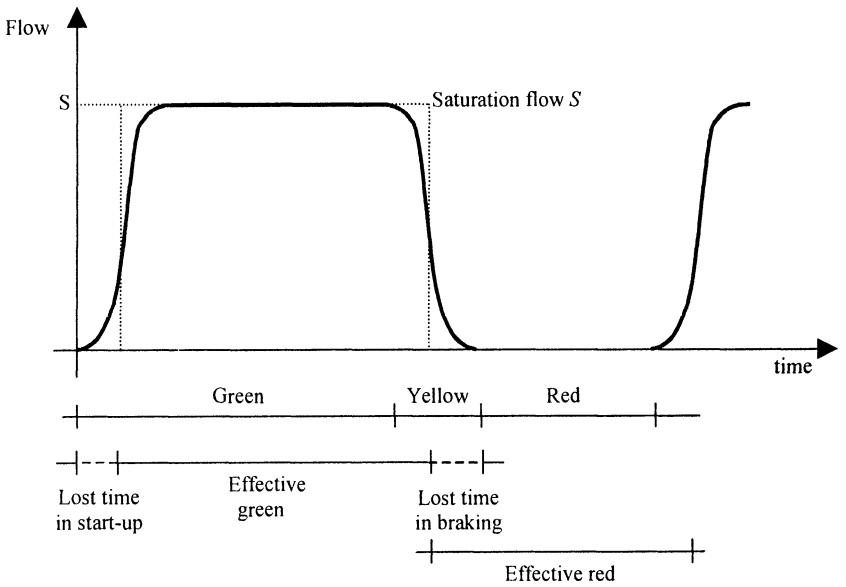


Fig. 2.3.8 Discharge flow from signal-controlled intersection in relation to cycle phases.

The saturation flow rate of an intersection can in principle be obtained through specific traffic surveys; in practice, however, empirical models based on average results are often used. The Highway Capacity Manual (HCM) describes one of the most popular methods. To apply this method, it is necessary to determine appropriate lane groups. A lane group is defined as one or more lanes of an intersection approach serving one or more traffic movements to which a single value of saturation flow, capacity and delay can be associated. Both the geometry of the intersection and the distribution of traffic movements are taken into account to segment the intersection into lane groups. In general, the smallest number of lane

groups that adequately describes the operation of the intersection is used. Fig. 2.3.9a shows some common lane group schemes suggested by the HCM.

The saturation flow rate of an intersection is computed from an "ideal" saturation flow rate, usually 1,900 equivalent passenger cars per hour of green time per lane (pcphgpl), adjusted for a variety of prevailing conditions that are not ideal. The method can be summarized by the following expression:

$$S = S_0 \cdot N \cdot F_w \cdot F_{HV} \cdot F_g \cdot F_p \cdot F_{bb} \cdot F_a \cdot F_{RT} \cdot F_{LT}$$

where:

S is the saturation flow rate for the specific lane group, expressed as a total for all lanes in the lane group under prevailing conditions, in vphg;
S_0 is the ideal saturation flow rate per lane, usually 1,900 pcphgpl;
N is the number of lanes in the lane group;
F_w is the adjustment factor for lane width (12 ft or 3.66 mt. lanes are standard);
F_{HV} is the adjustment factor for heavy vehicles in the traffic flow;
F_g is the adjustment factor for approach grade;
F_p is the adjustment factor for the existence of a parking lane adjacent to the lane group and the parking activity in that lane;
F_{bb} is the adjustment factor for the blocking effect of local buses that stop within the intersection area;
F_a is the adjustment factor for the area type;
F_{RT} is the adjustment factor for right turns in the lane group;
F_{LT} is the adjustment factor for left turns in the lane group.

The first six adjustment factors not connected with the type of turning maneuvers are reported in the Fig. 2.3.9b.

Once the approach capacity $Q_l = \mu S$ is known, the average waiting time tw_l can be calculated with several formulae.

One of the best known expressions is the *Webster* formula, proposed for an isolated intersection under the assumption of random (Poisson) arrivals and under-saturation conditions ($f_l/Q_l < 1$):

$$tw_l(f_l) = \frac{T_c(1-\mu_l)^2}{2(1-f_l/S_l)} + \frac{(f_l/Q_l)^2}{2f_l(1-f_l/Q_l)} - 0.65(Q_l/f_l^2)^{1/3}(f_l/Q_l)^{2+\mu} \qquad (2.3.10)$$

where:

T_c is the cycle length;
μ is the effective green to cycle length ratio for the lane group represented by link l;
Q_l is the capacity of the lane group represented by link l.

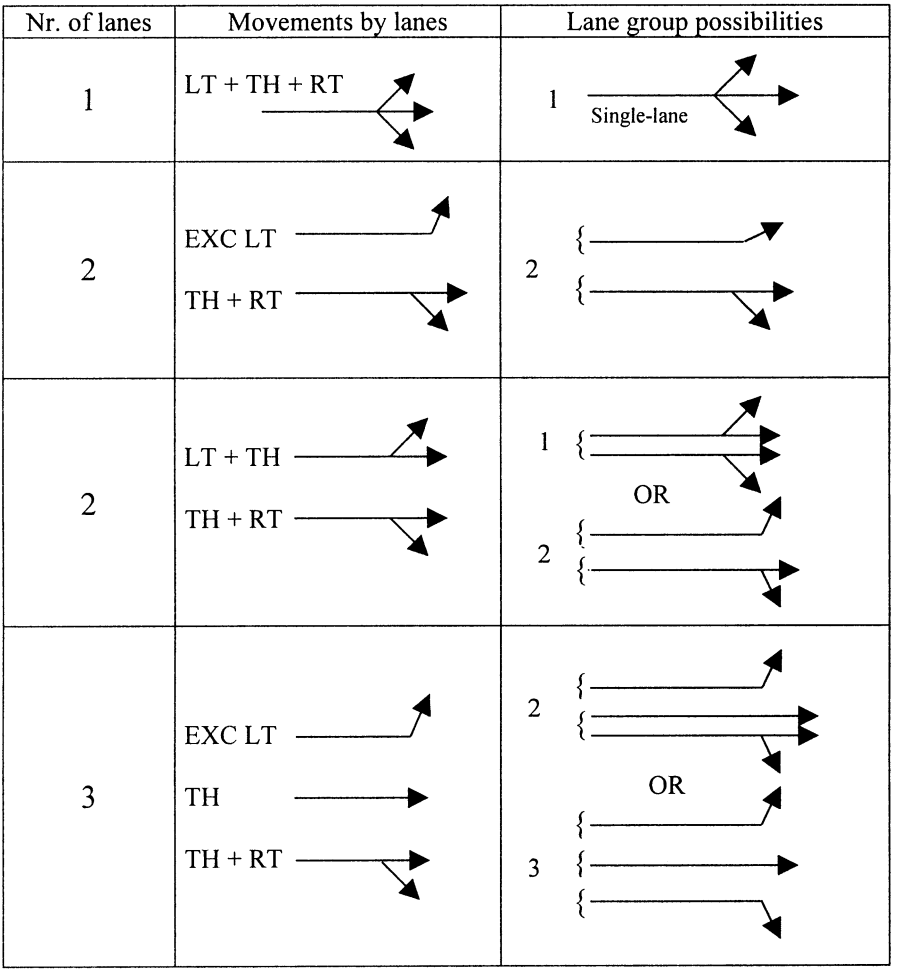


Fig. 2.3.9a Typical lane groups for the HCM method for calculating saturation flow.

ADJUSTMENT FACTOR FOR AVERAGE LANE WIDTH F_w									
Average lane width, W (FT)	8	9	10	11	12	13	14	15	16
F_w	0.867	0.900	0.933	0.967	1.000	1.033	0.067	1.100	1.133
ADJUSTMENT FACTOR FOR HEAVY VEHICLES F_{HV}									
Percentage of heavy vehicles (%)	0	2	4	6	8	10	15	20	
F_{HW}	1.000	0.980	0.962	0.943	0.926	0.909	0.870	0.833	
Percentage of heavy vehicles (%)	25	30	35	40	45	50	75	100	
F_{HW}	0.800	0.769	0.741	0.714	0.690	0.667	0.571	0.500	
ADJUSTMENT FACTOR FOR APPROACH GRADE F_g									
Grade (%)	-6	-4	-2	0	+2	+4	+6	+8	≥ 10
F_g	1.030	1.020	1.010	1.000	0.990	0.980	0.970	0.960	0.950

ADJUSTMENT FACTOR FOR PARKING F_p						
F_p	N. of parking maneuvers per hour					
N. of lanes in lane group	No parking	0	10	20	30	≥ 40
1	1.000	0.900	0.850	0.800	0.750	0.700
2	1.000	0.950	0.925	0.900	0.875	0.850
3 or more	1.000	0.967	0.950	0.933	0.917	0.900

ADJUSTMENT FACTOR FOR BUS BLOCKAGE F_{bb}					
F_{bb}	N. of buses stopping per hour				
N. of lanes in lane group	0	10	20	30	≥ 40
1	1.000	0.960	0.920	0.880	0.840
2	1.000	0.980	0.960	0.940	0.920
3 or more	1.000	0.987	0.973	0.960	0.947

ADJUSTMENT FACTOR FOR AREA TYPE F_a	
Type of area	F_a
CBD (Center Business District)	0.900
All other areas	1.000

Fig. 2.3.9b Adjustment factors in the HCM method for saturation flow.

The first term expresses the delay at zero flow, the second term expresses the delay due to congestion which tends to infinity if the flow tends to the capacity Q_l (see also section 2.A.3); the third term is a correction term obtained by simulation results.

Applying expression (2.3.10) for flows close to capacity, the delay are very large and, as observed previously, are both unrealistic and computationally problematic. Again, to correct for this, it is possible to apply the two heuristic methods described for equation (2.3.4). The first method applies equation (2.3.10) for values of f_l up to a percentage α of the capacity while for higher values a linear approximation of the function is used:

$$tw_l(f_l) = tw_l(\alpha Q_l) + \frac{d}{df} tw_l(f)\Big|_{f_l=\alpha Q_l} \cdot (f_l - \alpha Q_l) \quad f_l \geq \alpha Q_l \qquad (2.3.11)$$

Fig. 2.3.10 shows the diagram of the function (2.3.11) for a signalized intersection approach. The second method computes the over-saturation delay with a deterministic queuing model similar to that described in Fig. 2.3.7. The *Akcelik* formula for the calculation of the average delay at signalized intersections is derived from this approach:

$$tw_l(f_l) = \frac{0.5T_c(1-\mu_l)^2}{1-\mu_l X_l} \qquad X_l \le 0.50$$

$$tw_l(f_l) = \frac{0.5T_c(1-\mu_l)^2}{1-\mu_l X_l} + 900 \cdot T \cdot \left\{ X_l - 1 + \left[(X_l - 1)^2 + \frac{8(X_l - 0.5)}{\mu_l S_l T} \right]^{1/2} \right\} \quad 0.50 \le X_l \le 1$$

$$tw_l(f_l) = 0.5T_c(1-\mu_l) + 900 \cdot T \cdot \left\{ X_l - 1 + \left[(X_l - 1)^2 + \frac{8(X_l - 0.5)}{\mu_l S_l T} \right]^{1/2} \right\} \qquad X_l > 1$$

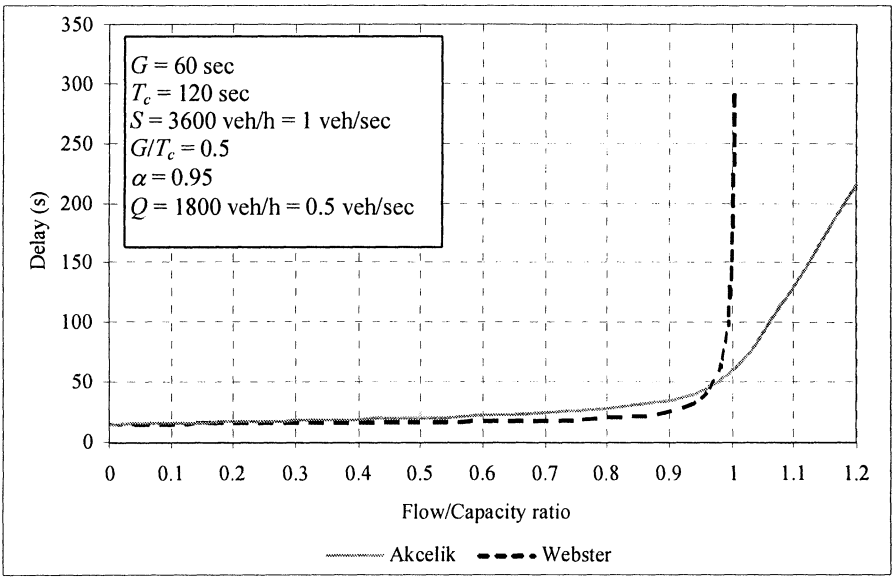
(2.3.12)

where $X_l = f_l/Q_l$ is the flow/capacity ratio, the times tw_l and T_c are expressed in seconds, S_l in pcph and T is the duration of over-saturation period in hours. Equation (2.3.12) is compared with the Webster formula in Fig. 2.3.10 for a value of T=0.5h.

Note that the application of the previous formulae for the calculation of saturation flows, capacities and average waiting times (delays) in case of multiple lane groups requires an "exploded" representation of the intersection with several links corresponding to the relevant lane groups and their maneuvers. For example, in the case of exclusive right-turn lane a single link can represent such a movement and the associated delay as in Fig. 2.3.3. Sometimes, to simplify the representation, fewer links than lane groups are used; in this case the total capacity of all lane groups is associated with the single link and the resulting delay is associated with the whole flow.

From a mathematical point of view the delay functions discussed so far are separable only if the traffic-signal regulation is such as to exclude interference between maneuvers represented by different links. For example this is the case for the 3-phase regulation scheme of a T-shaped intersection shown in Fig. 2.3.11. However, if the phases allow conflict points, e.g. left-turn from the opposite direction with through flows during the same phase, non-separable cost functions may be necessary since the left turn causes a reduction in the saturation flow and, thus, an increase on the delay for the through flows and vice-versa.

In general, if a single node represents the entire intersection, the effects of individual maneuvers and lane groups are impossible to distinguish and separable functions are adopted, with a single value of saturation flow, reduced to account for the interfering turns. When the exploded representation of the intersection is used, however, the delay functions of each maneuver might depend on the flow of another maneuver if the two maneuvers share some lanes and have the same green light phase (i.e. they belong to the same lane group). For example, for the four-arm, two-phase intersection in Fig. 2.3.11, the delay on the link corresponding to the maneuver (C-D) also depends on the flow on the link corresponding to the maneuver (C-A) if the two maneuvers share the same accumulation lanes and have the same green phase.



f	f/Q	Akcelik	Webster
0.00	0.00	15.00	15.00
0.10	0.20	16.67	16.87
0.20	0.40	18.75	19.26
0.25	0.50	20.00	20.77
0.30	0.60	21.93	22.61
0.40	0.80	27.95	28.45
0.50	1.00	60.00	
0.60	1.20	216.75	

Fig. 2.3.10 Waiting time functions at a signalized intersection.

Finally, in the case of more complex, flow actuated signal control systems, delay and cost functions are certainly not separable since the regulation parameters (length of the cycle T_c, green/cycle ratios μ) depend on the flows loading the various approaches converging at the intersection[2].

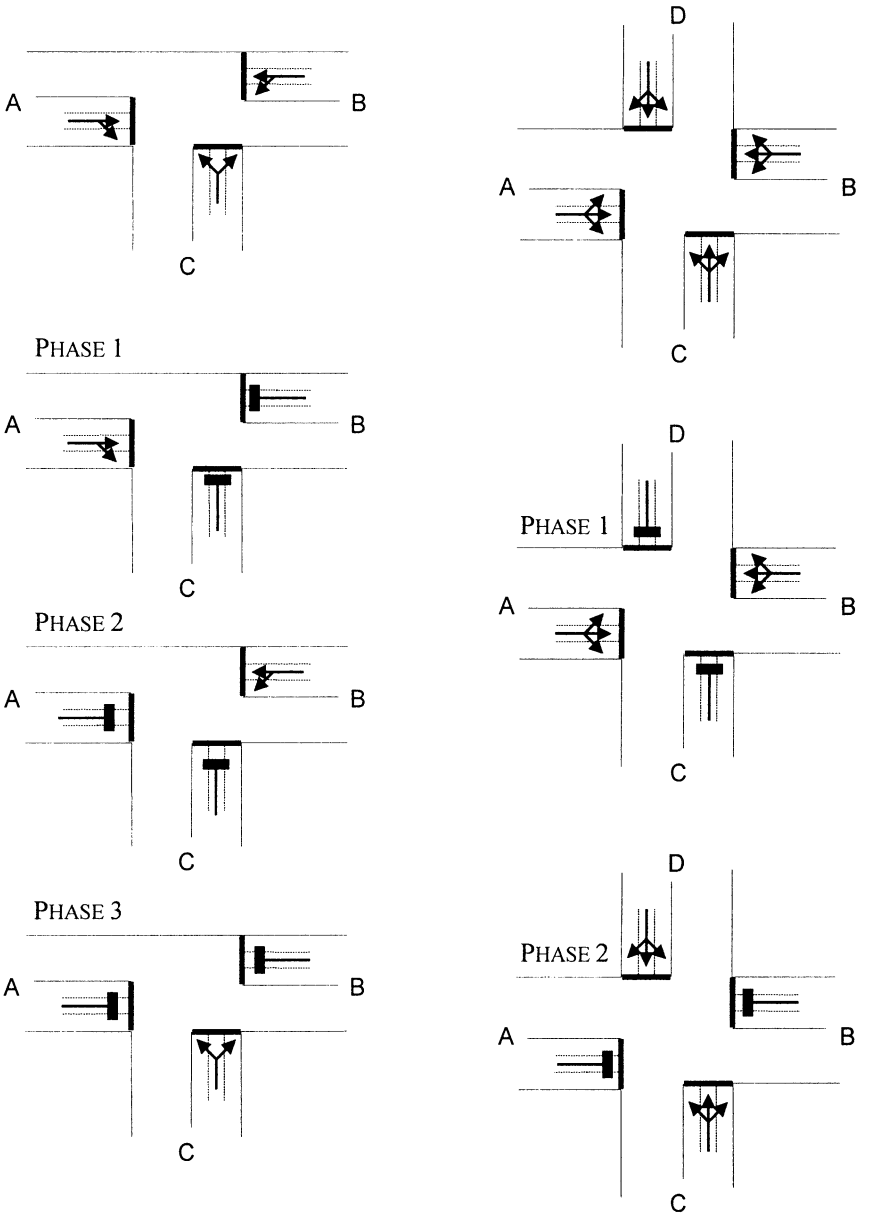


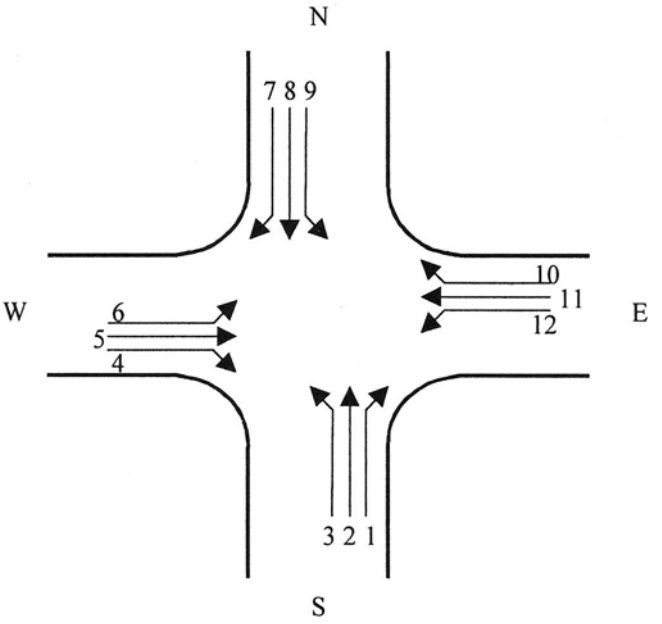
Fig. 2.3.11 Examples of traffic light phases for 3- and 4-arm intersections.

To complete the survey of the delay functions, *priority intersections,* i.e. intersections regulated by give-way rules rather than traffic lights, need to be considered. Empirical functions are often used to express average delays; these functions are non-separable in that right-of-way rules cause delays due to conflicts between flows. As an example, the delay corresponding to the maneuvers at a 4-arm intersection can be calculated by means of the following HCM function:

$$tw_l(f) = exp\{-0.2664 + 0.3967\ ln(f_{conf}) + 3.959\ A\ [ln(f_{conf}) - 6.92]\} \qquad (2.3.13)$$

where:

$tw_l(f)$ is the waiting time expressed in seconds;
f_{conf} is the total conflicting flow, which varies according to the maneuver as shown in Fig. 2.3.12;
$A = 1$ if $f_{conf} > 1062$ vehicles/hr, 0 otherwise.



Maneuver		Flows influencing the delay
Direction South-North, right-hand turn	1	1,2,3,5,9
Direction South-North, crossing	2	1,2,3,5,6,9,10,11,12
Direction South-North, left-hand turn	3	1,2,3,5,6,7,8,11,12

Fig. 2.3.12 Flows conflicts for computing delays in priority intersection.

f) *Parking links*

Monetary cost (fares) and search time are the most relevant performance attributes connected to links representing parking in a given area. In general these attributes are different for links representing different parking types (facilities). The more sophisticated models of search time take into account the congestion effect through the ratio between the average occupancy of the parking facilities of type p, represented by link l, and the parking capacity Q_l.

The average search time can be calculated through a model assuming that available parking spaces of type p are uniformly distributed along a circuit, possibly mixed with parking spaces of different types (e.g. free and priced parking). If the occupancy of a given parking type at the beginning and at the end of the reference period is inferior to the capacity, the following expression can be obtained:

$$ts_l(f_l) = \frac{L_p}{v_s} \frac{1}{occ_2(f_l) - occ_1} \cdot \frac{Q_{tot} \cdot (Q_p + 1)}{Q_p} \cdot \ln\left(\frac{1 + Q_p - occ_1}{1 + Q_p - occ_2(f_l)}\right) - \frac{(Q_{tot} - Q_p)}{Q_p} \tag{2.3.14}$$

where:

$ts_l(f_l)$ is the search time in minutes;
f_l is the flow on parking link l;
L_p is the average length of a parking space;
v_s is the average search speed for a free parking space;
occ_1 is the parking occupancy at the beginning of reference period;
occ_2 is the parking occupancy at the end of the reference period, depending on flow assigned to parking link and on the turn-over rate;
Q_p is the parking capacity of type p corresponding to link l;
Q_{tot} is the total capacity of all parking types mixed with type p in the zone.

In the case that one or both *occ* are above capacity, similar but formally more complicated formulas can be obtained. These expressions will not be reported here.

2.3.2. Supply models for scheduled service systems

Discontinuous and non-simultaneous transportation services can be accessed only at given points and are available only at given instants. Typical examples are scheduled services (buses, trains, airplanes, etc.), which can be used only between terminals (bus stops, stations, airports, etc.) and are available only at certain instants (departure times). Scheduled services can be represented by different supply models according to their characteristics and to the consequent assumptions on users' behavior (see section 4.2.5.). The approach followed in this chapter is based upon the modeling of *service lines*, i.e. a set of scheduled runs with equal characteristics.

This approach is consistent with the assumption of intra-period stationarity and with path choice behavior, typical of high frequency and irregular urban transit systems.

If service frequency is low and/or it is assumed that the users choose specific runs, it is necessary to represent the service with a different graph known as a *run graph* or *diachronic graph*. This is usually the case with extra-urban transportation services (airplanes, trains, etc.), which have low service frequencies and are largely punctual. In this case, however, the assumption of within-day stationarity does not hold. As a matter of fact, the supply characteristics often are not uniform within the reference period (arrival and departure times of single runs may be non-uniformly spaced). Furthermore, in order to simulate the traveler's behavior *desired departure* or *arrival times* should be introduced. For these reasons run-based supply models will be described in Chapter 6 dealing with intra-period dynamic systems.

2.3.2.1. Line-based graph models

If the scheduled services have high frequencies (e.g. one run every 5-15 minutes) and low regularity, it is usually assumed that the users do not choose an individual run, but rather a service line or a group of lines. A *service line* is a set of runs sharing the same terminals, the same intermediate stops and the same performance characteristics, as in the case of an urban bus or underground lines. In this case a *line graph* is typically used. In this graph, nodes correspond to stops, and more precisely to the relevant events occurring at the stops. *Access nodes* represent the arrival of the user at the stop, the *stop node*, or *diversion node*, represents the boarding of a vehicle, and the *line nodes* represent the arrival and departure of vehicles of a given line at a given stop. The links represent activities or phases of a trip: access trips between access nodes (*access links*), waiting at the stop (*waiting links*), boarding and alighting from the vehicles of a line (*boarding* and *alighting links*), the trip from one stop to another of the same line (*line links*), and vehicle dwelling at the stop (*dwelling links*).

Essentially, each stop is represented by a sub-graph such as the one shown in Fig. 2.3.13. The graph representing an entire public transportation system can be built by combining the *line graph* and the *access graph* through the stop sub-graphs. Access links may represent different access modes depending on the system modeled. In urban areas, they may represent pedestrian connections or, sometimes, undifferentiated "access modes" including local transit lines to the main network of bus and rail services. The line graph is completed by adding nodes and links allowing entry/exit from the centroids to the stops; in the urban context this usually occurs through pedestrian nodes and links or through road links connected to park-and-ride facilities (nodes).

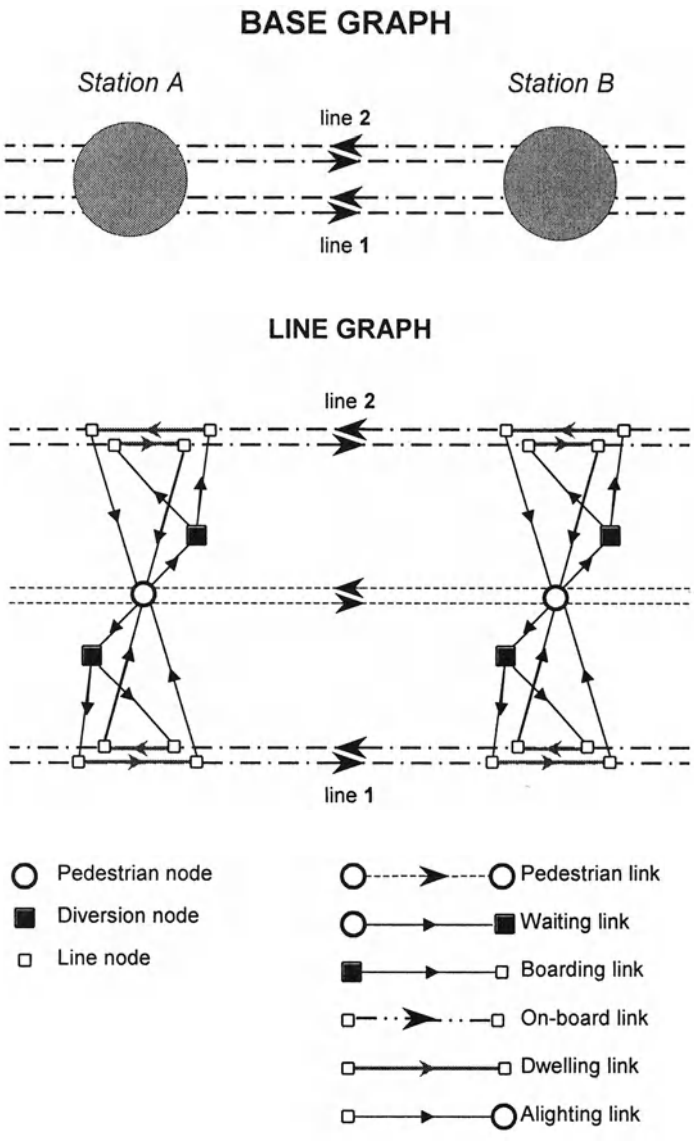


Fig. 2.3.13 Line-based graph for urban transit systems.

2.3.2.2. Performance and cost functions

The typical performance attributes used in line-based supply models are travel time components related to different trip phases and monetary costs. Travel times can be decomposed into on-board travel times, *Tb*, dwelling times at stops, *Td*, waiting times, *Tw*, boarding times, *Tbr*, alighting times, *Tal* and access/egress times, *Ta*, which may correspond to walking or driving time for urban transit networks. In general a single time component is associated to each link and the coefficients, β, homogenizing travel times into costs (disutilities) are different. In fact, several empirical studies have shown that waiting and walking times have coefficients two-three times larger than that of on board time for urban transit systems.

Performance functions used in many applications do not take into account congestion, at least with respect to flows of transit users, as it is assumed that services are designed with some extra capacity with respect to maximum user flows.

On-board travel time of a transit link can be obtained through a very simple expression:

$$Tb_l = \frac{L_l}{v_l(\boldsymbol{b}_l, \gamma_l)} \tag{2.3.15}$$

where vector $\boldsymbol{b}_l$ includes the relevant characteristics of the transit system represented by link l, while vector γ_l comprises a set of parameters. The average speed is strongly dependent on the type of right-of-way. For exclusive right-of-way systems, such as trains, the average speed, v_l, can be expressed as a function of the characteristics of the vehicles (weight, power, etc.), of the infrastructure (slope, radius of bends, etc.), of the circulation regulations on the physical section and the type of service represented. Relationships of this type can be deduced from mechanics for which specialized texts should be referred to. For partial right-of-way systems, such as surface buses, the average speed depends on the level of protection (e.g. reserved bus-lane) and the vehicle flows on the links corresponding to interfering movements. Performance functions of this type typically derive from descriptive models.

The *waiting time* is the average time that the users spend between their arrival at the stop/station and the arrival of the line (or lines) they board. Waiting time is usually expressed as a function of the line *frequency* φ_l, i.e. the average number of runs of line l in the reference period. When only one line is available the average waiting time, Tw_l, will depend on the regularity of vehicle arrivals and the pattern of users' arrivals to the stop. It can be shown that, under the assumption that users arrive at the stop according to a Poisson process with a constant arrival rate[3] (consistent with the within-day stationarity assumption), the average waiting time is:

$$Tw_l = \frac{\theta}{\varphi_l} \tag{2.3.16}$$

where θ is equal to 0.5 if the line is perfectly regular, i.e. the headways between successive vehicle arrivals are constant, it is equal to 1 if the line is "completely irregular", i.e. the headways between successive arrivals are distributed according to a negative exponential random variable; see Fig. 2.3.14.

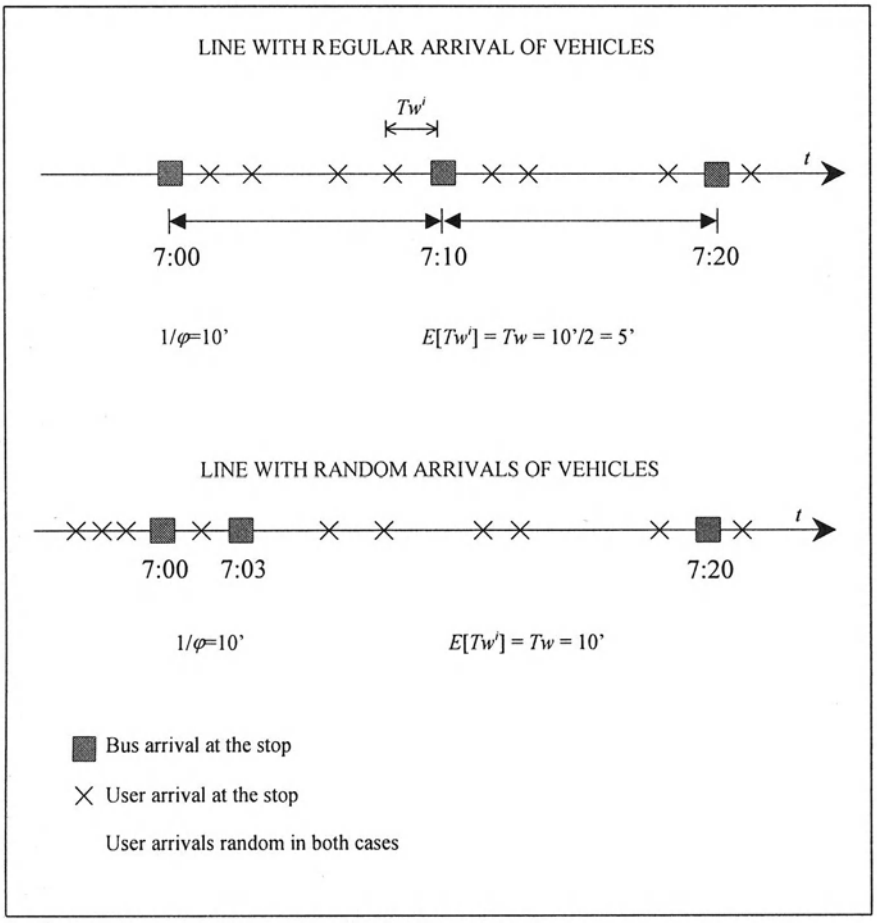


Fig. 2.3.14 Arrivals and waiting times at a bus stop.

In the case of several "*attractive lines*", i.e. when the user waits at a diversion node, m, for the first vehicle among those belonging to a set of lines, AL_m, the average waiting time can again be calculated with expression (2.3.16) by using the *cumulated frequency* Φ_m of the set of attractive lines[4]:

$$Tw_l = \frac{\theta}{\Phi_m} \quad \text{with} \quad \Phi_m = \sum_{l \in AL_m} \varphi_l \tag{2.3.17}$$

These expressions of average waiting times will be revisited in section 4.3.4.2 on the path choice models.

Access/egress times are also usually modeled through very simple performance functions analogous to expression (2.3.15):

$$Ta_l = \frac{L_l}{v_{al}(\boldsymbol{b}_l, \gamma_l)}$$

where v_{al} represent the average speed of the access/egress mode. Also in the case of pedestrian systems, it is possible to introduce congestion phenomena and correlate the generalized transportation cost with the pedestrian density on each section by using empirical expressions described in the literature.

More detailed performance models introduce congestion effects with respect to user flows both on travel times and on comfort performance attributes. An example of the first type of function is that relating the *dwelling time* at a stop, Td_l, to the user flows boarding and alighting the vehicles of each line:

$$Td_l = \gamma_1 + \gamma_2 \left(\frac{f_{al(l)} + f_{br(l)}}{Q_D} \right)^{\gamma_3} \tag{2.3.18}$$

where:

$f_{al(l)}$ is the users flow on the alighting link;
$f_{br(l)}$ is the users flow on the boarding link;
Q_D is the door capacity of the vehicle;
γ_1, γ_2, γ_3 are parameters of the function.

Another example is the function relating the average waiting time to the flow of users staying on board and those waiting to board a single line. This function takes into account the "refusal" probability, i.e. the probability that some users may not be able to get on the first arriving run of a given line because it is too crowded and have to wait longer for a subsequent one. In the case of a single attractive line l the waiting time function can be formally expressed as:

$$Tw_l = \frac{\theta}{\varphi_l} \left(\frac{f_{b(.)} + f_{w(.)}}{Q_l} \right) \tag{2.3.19}$$

where $\varphi_l(.)$ is the actual available frequency of line l, i.e. the average number of runs of the line for which there are available places. It depends on the ratio between the demand for places - sum of the user flow staying on board, $f_{b(.)}$, and the user flow willing to board, $f_{w(.)}$, - and the line capacity Q_l. This formula is valid only for $f_{b(.)} + f_{w(.)} > Q_l$.

Note that both performance functions (2.3.18) and (2.3.19) are non-separable, in that they depend on flows on links other that the one they refer to.

Discomfort functions relate the average riding discomfort on a given line section represented by link l, dc_l, to the ratio between the flow on the link (average number of users on board) and the available line capacity Q_l:

$$dc_l = \gamma_3 f_l + \gamma_4 \left(\frac{f_l}{Q_l} \right)^{\gamma_5} \tag{2.3.20}$$

where, as usual, γ_3, γ_4 and γ_5 are positive parameters, usually with γ_5 larger than one expressing more-than-linear effect of crowding.

2.A. Review of Traffic Flow Theory[°]

Traffic flow theory and related models simulate the effects of the interactions among vehicles simultaneously using a given transportation facility or service. The main scope of this review is to provide a theoretical background for the specification of performance functions. This appendix also provides basic concepts used in within-day dynamic supply models covered in Chapter 6. Many traffic flow models proposed in the literature are oriented more towards traffic operations rather than transportation system planning and design. Thus, a systematic analysis of traffic flow theory models is out of the scope of this book. In this appendix, notations are slightly different from common use in traffic flow theory to be consistent within this book.

For simplicity, models presented in this appendix will make reference to cars. However, most models can be applied to other types of users interacting while travelling along the same infrastructure such as trains, airplanes, and pedestrians. The models described below can be classified in two main groups by the type of phenomenon simulated. Models for *running links* simulate vehicle movements along a linear facility (e.g. a street), while models for *queuing links* simulate vehicles waiting to be served at a service station (e.g. a tollbooth).

2.A.1. Models for running links

Models for *running links* simulate the interaction among several users moving along the same transportation facility. These models can be derived from some simple results of traffic flow theory.

2.A.1.1. Fundamental variables

Several variables can be observed in a *traffic stream*, i.e. a sequence of cars moving along a road segment referred to as a link, l. In principle all variables should be related to link l, however, to simplify notation, the subscript l may be implied. The fundamental variables are the following, see Fig. 2.A.1:

τ the time at which the traffic is observed;
L_l the length of road segment corresponding to link l;
s a point along a link, or better its abscissa increasing (from a given origin, usually located at the beginning of the link) along traffic direction ($s \in [0, L_l]$);
i an index denoting an observed vehicle;
$v_i(s, \tau)$ the speed of vehicle i at time τ while traversing point (abscissa) s.

[°] Giulio Erberto Cantarella is co-author of this appendix.

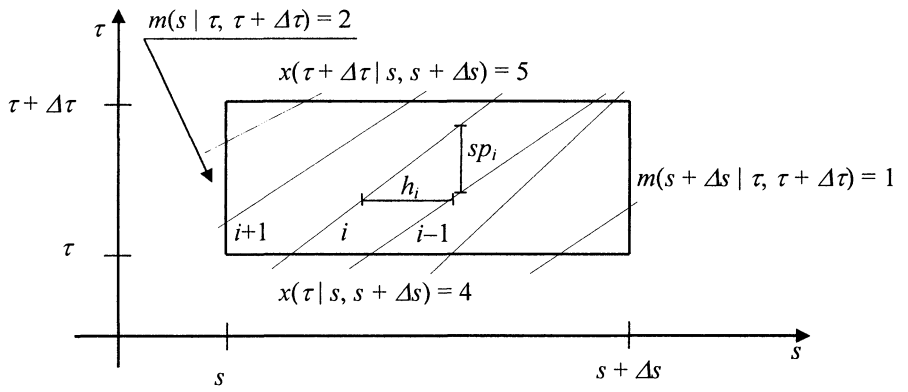


Fig. 2.A.1 Vehicle trajectories and traffic variables for running links.

For traffic observed at point s during the time interval $[\tau, \tau+\Delta\tau]$, several variables can be defined (see Fig. 2.A.1):

$h_i(s)$	the headway between vehicles i and $i-1$ crossing point s;
$m(s \mid \tau, \tau+\Delta\tau)$	the number of vehicles traversing point s during time interval $[\tau, \tau+\Delta\tau]$;
$\bar{h}(s)$	$= \Sigma_{i=1,\ldots,m}\, h_i(s) \,/\, m(s \mid \tau, \tau+\Delta\tau)$ the mean headway, among all vehicles crossing point s during time interval $[\tau, \tau+\Delta\tau]$;
$\bar{v}_t(s)$	$= \Sigma_{i=1,\ldots,m}\, v_i\,(s) \,/\, m(s \mid \tau, \tau+\Delta\tau)$ the time mean speed, among all vehicles crossing point s during time interval $[\tau, \tau+\Delta\tau]$.

Similarly, for traffic observed at time τ between points s and $s+\Delta s$, the following variables can be defined (see Fig. 2.A.1):

$sp_i(\tau)$	the spacing between vehicles i and $i-1$ at time τ;
$x(\tau \mid s, s+\Delta s)$	the number of vehicles at time τ between points s and $s+\Delta s$;
$\bar{sp}(\tau)$	$= \Sigma_{i=1,\ldots,x}\, sp_i(\tau) \,/\, x(\tau \mid s, s+\Delta s)$ the mean spacing, among all vehicles between points s and $s+\Delta s$ at time τ;
$\bar{v}_s(\tau)$	$= \Sigma_{i=1,\ldots,x}\, v_i \,/\, x(\tau \mid s, s+\Delta s)$ the space mean speed, among all vehicles between points s and $s+\Delta s$ at time τ.

During time interval $[\tau, \tau+\Delta\tau]$ between points s and $s+\Delta s$, a *general flow conservation equation* can be written:

$$\Delta x(s, s+\Delta s, \tau, \tau+\Delta\tau) + \Delta m(s, s+\Delta s, \tau, \tau+\Delta\tau) = \Delta z(s, s+\Delta s, \tau, \tau+\Delta\tau) \qquad (2.A.1)$$

where:

$\Delta x(s, s+\Delta s, \tau, \tau+\Delta\tau)$ $= x(\tau+\Delta\tau \mid s, s+\Delta s) - x(\tau \mid s, s+\Delta s)$ is the variation of the number of vehicles between points s and $s+\Delta s$ during $\Delta\tau$;

$\Delta m(s, s+\Delta s, \tau, \tau+\Delta\tau)$ $= m(s+\Delta s \mid \tau, \tau+\Delta\tau) - m(s \mid \tau, \tau+\Delta\tau)$ is the variation of the number of vehicles during time interval $[\tau, \tau+\Delta\tau]$ over space Δs;

$\Delta z(s, s+\Delta s, \tau, \tau+\Delta\tau)$ is the number of entering minus exiting vehicles (if any) during time interval $[\tau, \tau+\Delta\tau]$, due to entry/exit points (e.g. on/off ramps), between points s and $s+\Delta s$.

In the example of Fig. 2.A.1 there are no vehicles entering/exiting in the segment Δs, and Δx is equal to 1 while Δm is equal to -1.

With the observed quantities two relevant variables, *flow* and *density*, can be introduced:

$f(s \mid \tau, \tau+\Delta\tau) = m(s \mid \tau, \tau+\Delta\tau) / \Delta\tau$ is the flow of vehicles crossing point s during time interval $[\tau, \tau+\Delta\tau]$, measured in vehicles per unit of time;

$k(\tau \mid s, s+\Delta s) = x(\tau \mid s, s+\Delta s) / \Delta s$ is the density (or occupancy) between points s and $s+\Delta s$ at time τ, measured in vehicles per unit of length.

The flows at extremes of the road segment are denoted by special names and are represented by specific variables:

$u_l(\tau, \tau+\Delta\tau) = f_l(0 \mid \tau, \tau+\Delta\tau)$ the *inflow*, i.e. the flow entering link l during time interval $[\tau, \tau+\Delta\tau]$;

$w_l(\tau, \tau+\Delta\tau) = f_l(L_l \mid \tau, \tau+\Delta\tau)$ the *outflow*, i.e. the flow exiting link l during time interval $[\tau, \tau+\Delta\tau]$.

Flow and density are related to mean headway and mean spacing through the following relations:

$$f(s \mid \tau, \tau+\Delta\tau) \cong 1 / \bar{h}(s)$$

$$k(\tau \mid s, s+\Delta s) \cong 1 / \bar{sp}(\tau)$$

Note that if observations are perfectly synchronized with vehicles, the almost equality in the previous two equations becomes a proper equality.

Moreover, if the general flow conservation equation (2.A.1) is divided by $\Delta\tau$, the following equation is obtained:

$$\Delta x / \Delta\tau + \Delta f = \Delta e \qquad (2.A.2)$$

where:

$\Delta f(s, s+\Delta s, \tau, \tau+\Delta\tau) = \Delta m(s, s+\Delta s, \tau, \tau+\Delta\tau) / \Delta\tau$ is the variation of the flow over space;
$\Delta e(s, s+\Delta s, \tau, \tau+\Delta\tau) = \Delta z(s, s+\Delta s, \tau, \tau+\Delta\tau) / \Delta\tau$ is the (net) entering/exiting flow.

At the same time, dividing by Δs, the general flow conservation equation (2.A.1) becomes:

$$\Delta k / \Delta\tau + \Delta f / \Delta s = \Delta e / \Delta s \qquad (2.A.3)$$

where:

$\Delta k (s, s+\Delta s, \tau, \tau+\Delta\tau) = \Delta x(s, s+\Delta s, \tau, \tau+\Delta\tau) / \Delta s$ is the variation of the density over time.

2.A.1.2. Stationary models

Traffic flow is called *stationary* during a time interval $[\tau, \tau+\Delta\tau]$ between points s and $s+\Delta s$ if flow is (on average) independent of point s, and density is independent of time τ:

$$f(s \mid \tau, \tau+\Delta\tau) = f = \bar{u} = \bar{w}$$

$$k(\tau \mid s, s+\Delta s) = k$$

In this case, the time mean speed is independent of location and space mean speed is independent of time:

$$\bar{v}_\tau(s) = \bar{v}_\tau$$

$$\bar{v}_s(\tau) = \bar{v}_s$$

Let $x = k \cdot \Delta s$ be the number of vehicles between points s and $s+\Delta s$ at any time during the interval $[\tau, \tau+\Delta\tau]$ and let $\bar{v}_s$ be the space mean speed of these vehicles. The vehicle at point s at time τ, on average, will reach point $s+\Delta s$ at time $\tau+\Delta s / \bar{v}_s$. Thus, on average, all m vehicles can cross point $s+\Delta s$ in a time $\Delta\tau' = \Delta s / \bar{v}_s$; the number of vehicles crossing point $s+\Delta s$ during time $\Delta\tau'$ is $f \cdot \Delta\tau'$. Therefore, the average number of vehicles crossing point $s+\Delta s$ must equal the average number (independent of time due to stationarity) of vehicles in the segment $[s, s+\Delta s]$ (see Fig. 2.A.2):

$$k \Delta s = f \Delta s / \bar{v}_s$$

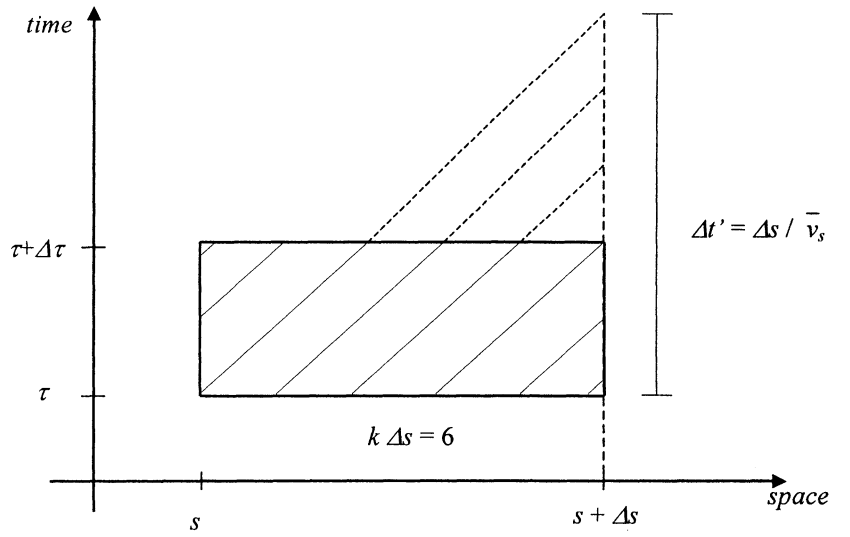


Fig. 2.A.2 Vehicle trajectories and traffic variables for stationary (deterministic) flows on running links.

Hence, under stationary conditions flow, density and space mean speed must satisfy the *stationary flow conservation equation*:

$$f = k\,v \tag{2.A.4}$$

where $v = \bar{v}_s$ is the space mean speed, simply called speed for further analysis of stationary conditions[5].

Multiple vehicles using the same facility may interact with each other and the effect of their interaction will increase with the number of vehicles. This phenomenon, called *congestion*, occurs in most transportation systems, generally worsening the overall performances of the facility, such as the mean speed or the travel time, since a vehicle may not be able to move at the desired speed. Stochastic models can be used to estimate the probability that a vehicle may be slowed down by another vehicle, as a function of the flow or density and desired speed distribution among vehicles. These considerations are particularly relevant when studying systems with scheduled services where congestion arises from out-of-schedule vehicles without the possibility of overtaking.

Congested systems with continuous service can be also modeled through (aggregate) deterministic models. In fact, under stationary conditions, aggregate

relationships may be observed between any pair of variables: flow, density and speed:

$$v = v(f) \tag{2.A.5}$$

$$v = V(k) \tag{2.A.6}$$

$$f = f(k) \tag{2.A.7}$$

This approach is less effective for systems with scheduled services, where the flow is generally small and averages are less meaningful.

Generally, observed values are rather dispersed (see Fig. 2.A.3 for a speed-flow relationship) and several models may fit observed values.

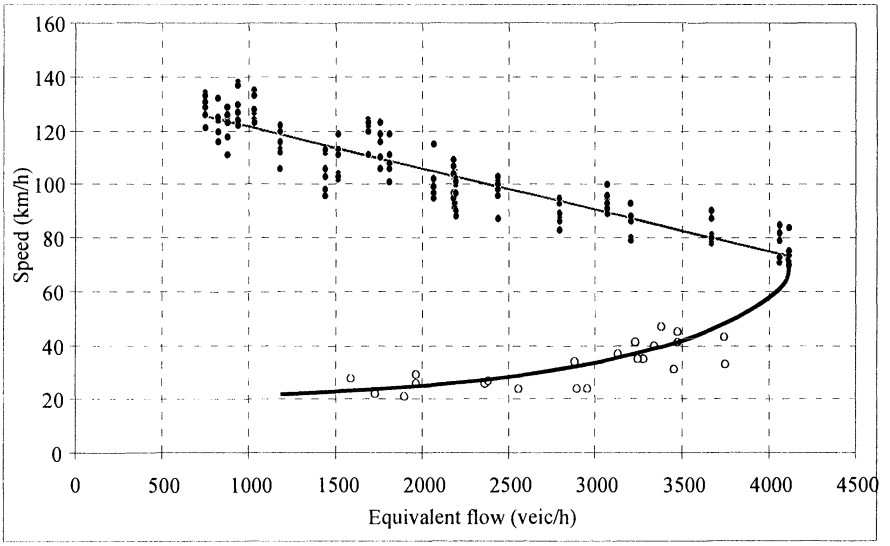


Fig. 2.A.3 Relationship between speed and flow.

The general form of relationships (2.A.5), (2.A.6) and (2.A.7) is illustrated in Fig. 2.A.3, also known as the *fundamental diagram of traffic flow*. This diagram shows that flow may be zero under two conditions: when density is zero (no vehicles on the road) or when speed is zero (vehicles are not moving). In the first case the speed assumes the theoretical maximum value, *free-flow speed*, v_0, while in the second the density assumes the theoretical maximum value, *jam density*, k_{jam}. Therefore, a traffic stream may be modeled through a *partially compressible fluid*, i.e. a fluid that can be compressed up to a maximum value.

The peak of the *speed-flow* (and *density-flow*) curve occurs at the theoretical maximum flow, *capacity*, Q, of the facility; the corresponding speed v_c and density k_c are referred to as the *critical speed* and the *critical density*.

Thus any value of flow (except the capacity) may occur under two different conditions: low speed and high density and high speed and low density.

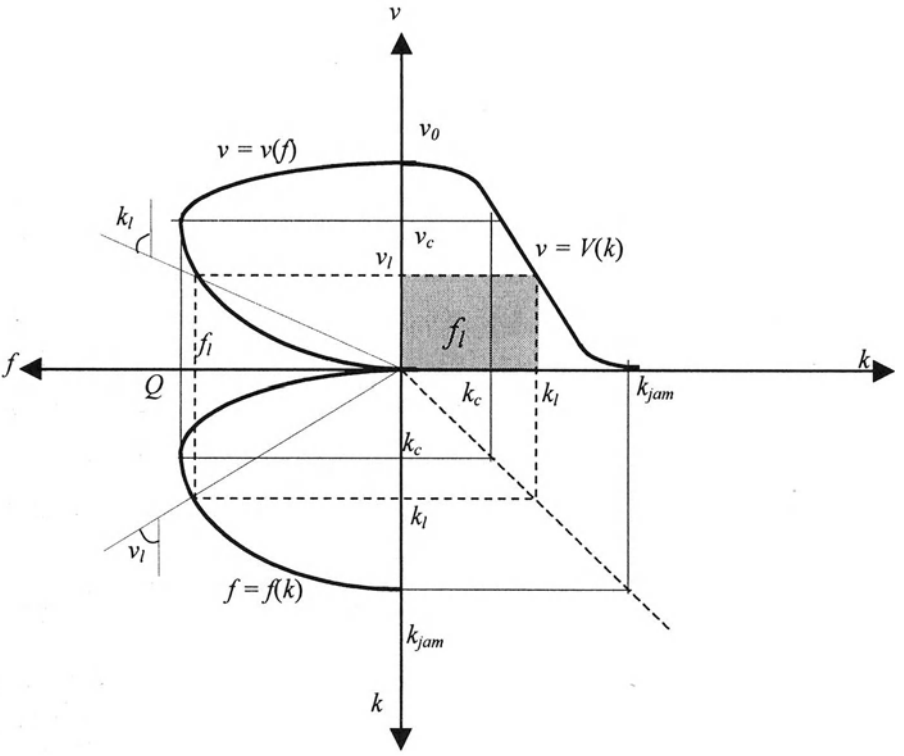


Fig. 2.A.4 Fundamental diagram of traffic flow.

The first condition represents an unstable state for the traffic stream, where any increase of density will cause a decrease in speed and thus in flow. This action produces another increase of density and so on until traffic becomes jammed. Conversely, the second condition is a stable state since any increase in density will cause a decrease in speed and an increase in flow.

At capacity (or at critical speed or at critical density), the traffic stream is relatively unstable; in fact, if the density increases, the speed will decrease and traffic will become unstable.

These results show that flow cannot be used as the unique parameter describing the state of a traffic stream; speed and density, instead, can univocally identify the prevailing traffic condition. For this reason the relation $v = V(k)$ is preferred to study traffic stream characteristics.

In literature several authors have proposed mathematical formulations for the fundamental diagram, based on *single regime* or *multi regime* functions. An example of single regime function is the Greenshield's linear model:

$$V(k) = v_0 (1 - k/k_{jam})$$

An example of a multi regime function is the Greenberg's model:

$$V(k) = a_1 \ln (a_2/k) \qquad \text{for } k > k_{min}$$

$$V(k) = a_1 \ln (a_2/k_{min}) \qquad \text{for } k \leq k_{min}$$

where a_1, a_2 and $k_{min} \leq k_{jam}$ are constants to be calibrated.

Starting from the speed-density relationship, the flow-density relationships $f(k)$ can be obtained as:

$$f(k) = V(k)\, k$$

The Greenshield's linear model yields:

$$f(k) = v_0 (k - k^2 / k_{jam})$$

In this case the capacity is given by:

$$Q = v_0 k_{jam}/4$$

Moreover the flow-speed relationship can be obtained by introducing the inverse speed-density relationship:

$$k = V^{-1}(v)$$

thus

$$f(v) = V(k = V^{-1}(v)) \cdot V^{-1}(v) = v \cdot V^{-1}(v)$$

For example, the Greenshield's linear model yields:

$$V^{-1}(v) = k_{jam} (1 - v/v_0)$$

thus

$$f(v) = k_{jam} (v - v^2/v_0)$$

Generally, the flow-speed relationship cannot be inverted since for each value of speed in the range 0 and v_0 two values of flow exist, corresponding to the stable and unstable regimes. However, for the stable regime only (or the unstable one), an inverse relationship speed-flow can be obtained:

$$v = v(f)$$

For example, Greenshield's linear model yields:

$$v_{stable}(f) = \frac{v_0}{2}\left(1 + \sqrt{1 - 4f/(v_0 k_{jam})}\right) = \frac{v_0}{2}\left(1 + \sqrt{1 - 4f/Q}\right)$$

$$v_{unstable}(f) = \frac{v_0}{2}\left(1 - \sqrt{1 - 4f/Q}\right)$$

Starting from the (stable regime) speed-flow relation, the (stable regime) travel time of a running link l can be calculated as a function of flow:

$$tr_l = L_l / v_l(f_l) \qquad (2.A.8)$$

where:

tr_l is the running time on link l;
f_l is the flow on link l;
L_l is the length of the running link l;
v_l is the mean speed on the link l assuming a stable regime.

Alternatively, travel time on a link can be computed as a function of flow and free-flow speed, without an explicit speed-flow relation. Two examples of such travel time functions are (see Fig. 2.A.5):

polynomial (e.g. BPR)

$$tr_l = (L_l/v_{0\,l})\,(1 + \alpha(f_l/Q_l)^{\beta})$$

hyperbolic (e.g. Davidson)

$$\begin{cases} tr_l = (L_l/v_{0\,l})(1 + \gamma f_l / (Q_l - f_l)) & \text{for} \quad f_l \le \delta Q_l \\ tr_l = \text{tangent approximation} & \text{for} \quad f_l > \delta Q_l \end{cases}$$

with $\delta < 1$ and Q_l = link capacity.

In this last case the tangent approximation is necessary since tr_l tends to ∞ for f_l going to Q_l (compare with section 2.3.1.2). This condition is unrealistic because the over-saturated period has a finite duration.

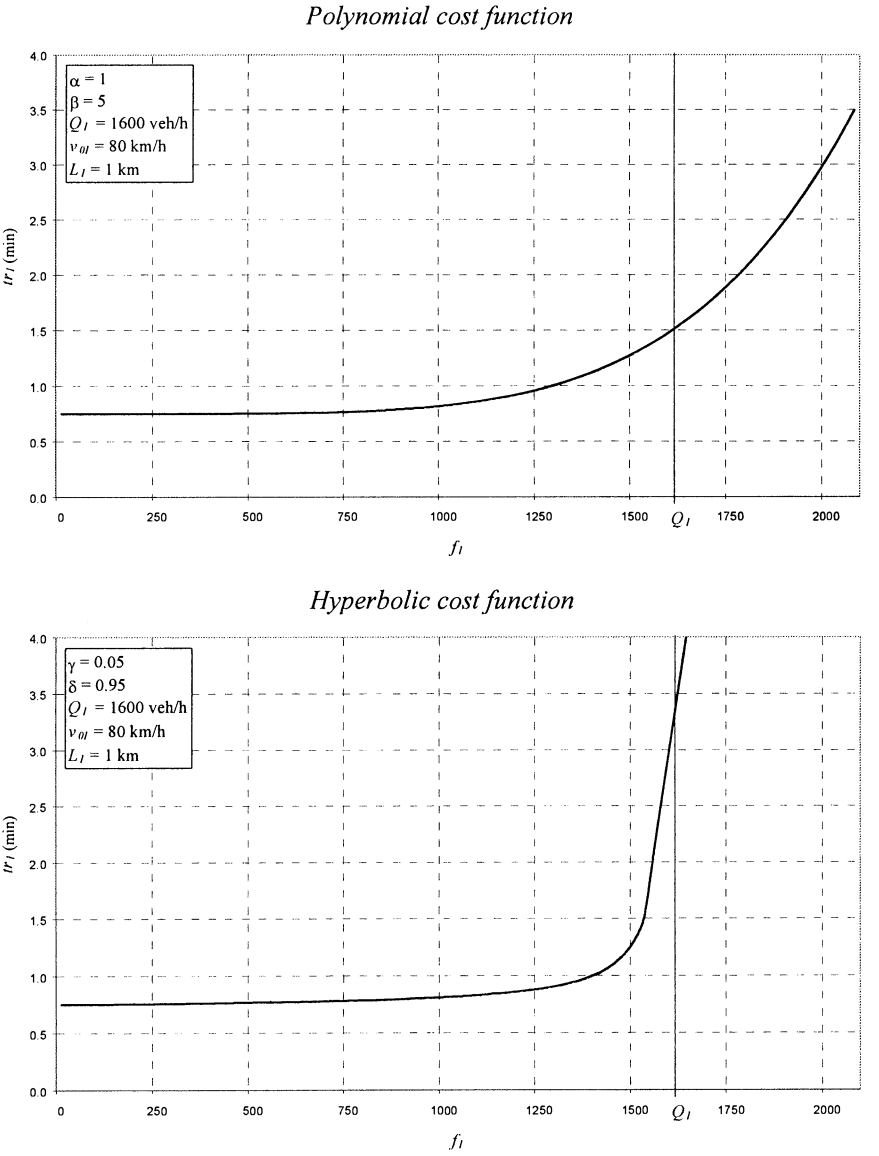


Fig. 2.A.5 Travel time flow functions.

2.A.1.3. Non-stationary models

Non-stationary models simulate explicitly variations of the main variables over space and time. For this reason they are also referred to as *dynamic traffic models*. Non-stationary models are mainly used in the context of within-day dynamic supply models described in Chapter 6. They are introduced in this appendix to give a complete overview of traffic flow theory. These models can be classified in three main classes by performance functions and flow representation (see Fig. 2.A.6):

- *Macroscopic models*: traffic is represented continuously following the fluid approximation (described in details below); individual trajectories are not explicitly traced. Aggregate performance measures are calculated using relations derived from stationary models.
- *Mesoscopic models*: traffic is represented discretely (vehicles or groups of vehicles); individual trajectories can be explicitly traced. Aggregate performance measures are calculated as for macroscopic models.
- *Microscopic models*: traffic is represented discretely (single vehicles); individual trajectories can be explicitly traced as for mesoscopic models. Disaggregate performance measures are calculated based on explicit modeling of driver behavior.

		Performance functions	
		Aggregate	*Disaggregate*
Flow Representation	*Continuous*	MACROSCOPIC	-
	Discrete	MESOSCOPIC	MICROSCOPIC

Fig. 2.A.6 Classification of non-stationary traffic models.

Microscopic models simulate the journey of each single vehicle through explicit driving behavior models of speed adjustment (e.g. car following, lane changing, overtaking, gap-acceptance, etc.). Such models can be solved only by event-based or time-based simulation techniques. Microscopic models provide very detailed traffic simulation on a small scale; yet require a significant amount of data and effort for specification and calibration. For these reasons microscopic models are used primarily for traffic operations rather than transportation planning and will not be further analyzed.

In mesoscopic models all vehicles on a link have the same speed, generally depending on density, varying over time. This type of model will be analyzed in Chapter 6 on dynamic traffic assignment.

Macroscopic models, described below, are based on *fluid approximation*; i.e. a traffic stream is represented through a partly compressible fluid, made up by infinitesimal particles. This fluid is described by point variables; following this assumption, flow and density are considered as function of point s and time τ:

$$f = f(s, \tau)$$

$$k = k(s, \tau)$$

These functions have only a mathematical interpretation since they cannot be observed in a discrete phenomenon; for simplicity, the same notation used for observed variables will be adopted. Previously introduced observed values are related to these functions through the following relations:

$$\int_{\tau}^{\tau+\Delta\tau} f(s, z)\, dz = m(s \mid \tau, \tau+\Delta\tau) \qquad (2.A.9)$$

$$\int_{s}^{s+\Delta s} f(z, \tau)\, dz = x(\tau \mid s, s+\Delta s) \qquad (2.A.10)$$

These relations represent the mono-dimensional fluid approximation. According to this analogy, speed may only vary with point s (and possibly time τ) but not along any direction orthogonal to axis s. Thus, a particle may not overtake any particle ahead (this condition is also called "first-in-first-out", or FIFO, rule) and all movements are parallel to axis s.

Macroscopic models are generally based on *conservation differential equations*, which can be specified through two different approaches. *Space discrete models* analyze traffic on a *link* base. Let

L_l be the length of link l;
$x_l(\tau)$ be the number of equivalent vehicles[6] on link l at time τ;
$k_l(\tau)$ be the (space) mean density on link l at time τ;
$u_l(\tau)$ be the flow entering link l at time τ;
$w_l(\tau)$ be the flow exiting from link l at time τ;
$t^f_l(\tau)$ be the (running) travel time of link l arriving at time τ.

For each link l the following *link flow conservation equation* holds:

$$\partial x_l(\tau) / \partial \tau = u_l(\tau) - w_l(\tau) \qquad (2.A.11)$$

Notice that (2.A.11) is equivalent to (2.A.2) in the observed variables.

The travel time on link l for a vehicle arriving at time τ can be expressed through (stationary) speed-density functions with respect to the number of vehicles x_l at time τ as:

$$t^f_l(\tau) = L_l / V_l(k_l = x_l(\tau)/L_l) \qquad (2.A.12)$$

where $V_l(k_l)$ is an empirical performance (speed-density) function analogous to those described for stationary models. Another relation can be added, expressing the mono-dimensional fluid conditions:

$$w_l(\tau + t^f_l(\tau)) = u_l(\tau) / (1 + \partial t^f_l(\tau) / \partial \tau) \qquad (2.A.13)$$

Thus, for each link three equations can be written in the four variables: $u_l(\tau)$, $w_l(\tau)$, $x_l(\tau)$ and $t^f_l(\tau)$. Once the entering flow is given, the others can be obtained. Hence the whole model is made up by equations (2.A.11), (2.A.12) and (2.A.13) for each link together with *node consistency equations*, that assure flow conservation at each node (for each O-D pair and/or user class). These last equations can be written as:

$$\sum_{l \in FS(n)} u_l(\tau) - \sum_{l \in BS(n)} w_l(\tau) = \begin{cases} 0 & \text{if } n \text{ is not a centroid} \\ G_n(\tau) - A_n(\tau) & \text{if } n \text{ is a centroid} \end{cases}$$

where:

$FS(n)$ is the set of links belonging to the forward star of node n;
$BS(n)$ is the set of links belonging to the backward star of node n;
$G_n(\tau)$ is the flow generated by the centroid node n;
$A_n(\tau)$ is the flow absorbed by the centroid node n.

Space continuous models analyze the traffic on a *point* base; they are based on the following *flow conservation equations* (equivalent to eqn (2.A.3) and (2.A.4) among observed variables):

$$\partial k(s,\tau) / \partial \tau + \partial f(s,\tau) / \partial s = 0 \qquad (2.A.14)$$

$$f(s,\tau) = k(s,\tau)\, v(s,\tau) \qquad (2.A.15)$$

assuming that no exit or entry occur at point s.

For each link a performance function should also be included. In *first order models* this function is directly derived from (stationary) speed-density functions:

$$v(s,\tau) = V(k(s,\tau))$$

For instance, using the Greenshield's relation, the so-called LWR model is obtained:

$$v(s,\tau) = v_0(1 - k(s,\tau) / k_{jam})$$

In *second order models* the performance function is specified through acceleration equations. For example, the Payne model is specified by adding the following equations:

$$\mathrm{d}v/\mathrm{d}\tau = \partial v/\partial \tau + v(s,\tau)\, \partial v/\partial s = (v(k(s,\tau)) - v(s,\tau))/\tau_{REC} - (\alpha_{ANT}/\tau_{REC})\, \partial k/\partial s$$

where τ_{REC} and α_{ANT} are parameters to be calibrated.

Also for space continuous models, consistency conditions must be written for any node.

Both space discrete and space continuous macroscopic models, specified by *differential equations*, can be solved by *finite difference approximation* methods based on time and/or space discretization. Thus, from the application point of view, a link segmentation to solve discrete space models gives results similar to a space difference formulation of a continuous space model. The need to satisfy mono-dimensional fluid approximation (FIFO rule) may require additional equations; thus, a whole finite difference macroscopic model can yield results similar to a mesoscopic one. These concepts will be further explored in section 6.2.

2.A.2. Models for queuing links

Models for *queuing links* simulate the interactions among several users waiting to receive a service at a given location (e.g. signalized intersections, bottlenecks, toll-booths, etc.). The location of the service is called the *server* and a *delay* is usually associated with the service activity represented by a queuing link. Queuing models can be studied following an approach similar to that used for running links, even though this similarity is not always recognized in the two specific literatures. The notation adopted in this section will underline this similarity.

2.A.2.1. Fundamental variables

The following focuses on delays related to a queue of users waiting for service upstream of the server. The main variables that describe queuing phenomena are:

τ	time at which the system is observed;
$x(\tau)$	number of users waiting to exit (queue length) at time τ;
$m_{IN}(\tau, \tau+\Delta\tau)$	number of users joining the queue during $[\tau, \tau+\Delta\tau]$;
$m_{OUT}(\tau, \tau+\Delta\tau)$	number of users leaving the queue during $[\tau, \tau+\Delta\tau]$;
$u(\tau, \tau+\Delta\tau)$	$= m_{IN}(\tau, \tau+\Delta\tau)/\Delta\tau$ arrival (entering) flow during $[\tau, \tau+\Delta\tau]$;
$w(\tau, \tau+\Delta\tau)$	$= m_{OUT}(\tau, \tau+\Delta\tau)/\Delta\tau$ exiting flow during $[\tau, \tau+\Delta\tau]$;
τ_i	arrival time of vehicle i;
h^{IN}_i	headway between arrival times τ_i and τ_{i-1};
h^{IN}	$(\tau, \Delta\tau)$ average arrival headway;
$T_{s\,i}$	service time of vehicle i;
$T_s(\tau, \Delta\tau)$	average service time;
tw_i	total waiting time (pure waiting plus service time) of vehicle i;
$tw\,(\tau, \tau+\Delta\tau)$	average waiting time;
$Q(\tau, \Delta\tau)$	$= 1/T_s(\tau, \Delta\tau)$ capacity or maximum exit flow, assumed constant during $[\tau, \tau+\Delta\tau]$ for simplicity's sake (otherwise $\Delta\tau$ can be redefined).

The exiting flow and the capacity are correlated by the capacity constraint:

$$w \leq Q$$

Notice that the main difference with the basic variables of running links is that space (s, Δs) is no longer explicitly referred to since it is irrelevant. Some of the above variables are shown in Fig. 2.A.7.

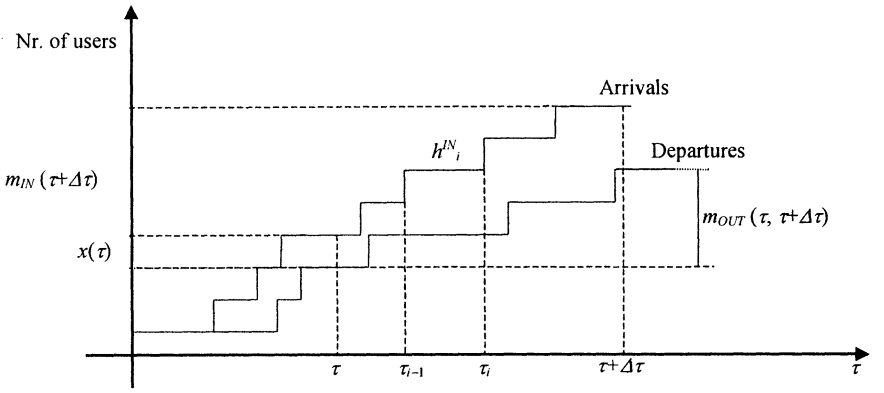


Fig. 2.A.7 Fundamental variables for queuing systems.

The models simulating queuing phenomena can be *deterministic* or *stochastic* depending on the assumptions for vehicle arrivals and service times, i.e. whether h_i and T_{si} are modeled as deterministic or random variables.

2.A.2.2. Deterministic models

Deterministic models are based on the assumptions that arrival and departure times are deterministic variables. In spite of the discrete nature of the phenomenon, deterministic queuing systems often are modeled and represented as continuous systems similarly to the fluid approximation of traffic flows (see Fig. 2.A.8). Flow conservation equations (2.A.1) and (2.A.2) introduced for running links still hold, leading to:

$$x(\tau) + m_{IN}(\tau,\ \tau+\Delta\tau) = m_{OUT}(\tau,\ \tau+\Delta\tau) + x(\tau+\Delta\tau) \qquad (2.A.16)$$

Dividing equation (2.A.16) for $\Delta\tau$ and remembering the definitions of entering and exiting flows, u and w, the flow conservation equation can be expressed also as:

$$\Delta x / \Delta\tau + [w(\tau,\ \tau+\Delta\tau) - u(\tau,\ \tau+\Delta\tau)] = 0 \qquad (2.A.17)$$

Taking the limit for $\Delta\tau \rightarrow 0$ we get:

$$\frac{\mathrm{d}x(\tau)}{\mathrm{d}t} = u(\tau) - w(\tau)$$

Deterministic queuing systems can also be analyzed through the cumulative number of users that have arrived at the *server* by time τ, and the cumulative number of users that have departed from the *server* (leaving the queue) at time τ, as expressed by two functions named *arrival curve*, $A(\tau)$, and *departure curve*, $D(\tau) \leq A(\tau)$, respectively, see Fig. 2.A.9.

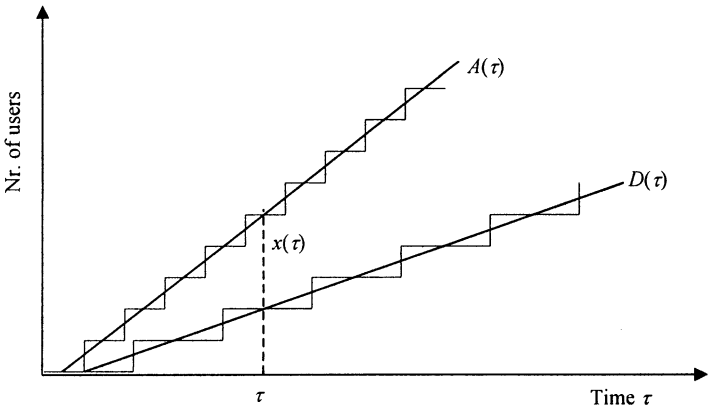


Fig. 2.A.8 Continuous approximation of stationary deterministic queuing systems.

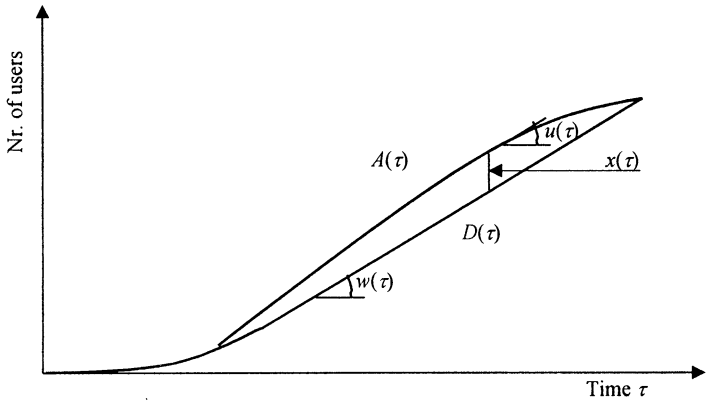


Fig. 2.A.9 Cumulative arrival and departure curves.

Queue length $x(\tau)$ at any time τ is given by:

$$x(\tau) = A(\tau) - D(\tau) \tag{2.A.18}$$

provided that the queue at time 0 is given by $x(0) = A(0) \geq 0$ with $D(0) = 0$.

The arrival and departure functions are linked to entering and exiting users by the following relationships:

$$m_{IN}(\tau, \tau+\Delta\tau) = A(\tau+\Delta\tau) - A(\tau) \tag{2.A.19}$$

$$m_{OUT}(\tau, \tau+\Delta\tau) = D(\tau+\Delta\tau) - D(\tau) \tag{2.A.20}$$

The *flow conservation equation* (2.A.16) can also be obtained by subtracting member by member the relationships (2.A.19) and (2.A.20) and taking into account equation (2.A.18).

The above equations (2.A.19) and (2.A.20) can be reformulated in terms of flow variables by dividing for $\Delta\tau$ and taking the limit for $\Delta\tau \to 0$, yielding (see Fig. 2.A.9):

$$u(\tau) = \frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau}$$

$$w(\tau) = \frac{\mathrm{d}D(\tau)}{\mathrm{d}\tau}$$

If during time interval $(\tau_0, \tau_0+\Delta\tau)$ the entering flow is constant over time, $u(\tau) = \bar{u}$, then the queuing system is named *stationary* (see Fig. 2.A.9) and the arrival function $A(\tau)$ is linear with slope given by $\bar{u}$:

$$A(\tau) = A(\tau_0) + \bar{u} \cdot (\tau - \tau_0) \qquad \tau \in [\tau_0, \tau_0+\Delta\tau]$$

The exit flow may be equal to the entering flow, $\bar{u}$, or to the capacity, Q, as described below.

In stationary queuing models used on transportation networks, the inflow $\bar{u}$ can be substituted with the flow f_l of the link representing the queuing system. Thus, in section 2.3.1.2 queuing delays are expressed as a function of f_l rather then $\bar{u}$ as in the following.

a) *Under-saturation*

When the arrival flow is less than capacity ($\bar{u} < Q$) the system is *under-saturated.* In this case, if there is a queue at time τ_0, its length decreases with time and vanishes after a time $\Delta\tau_0$ defined as (see Fig. 2.A.10):

$$\Delta\tau_0 = x(\tau_0)/(Q - \bar{u}) \tag{2.A.21}$$

Before time $\tau_0 + \Delta\tau_0$, the queue length is linearly decreasing with τ and the exiting flow $\bar{w}$ is equal to capacity:

$$x(\tau) = x(\tau_0) - (Q - \bar{u})(\tau - \tau_0) \tag{2.A.22}$$

$$\bar{w} = Q$$

$$D(\tau) = D(\tau_0) + Q(\tau - \tau_0)$$

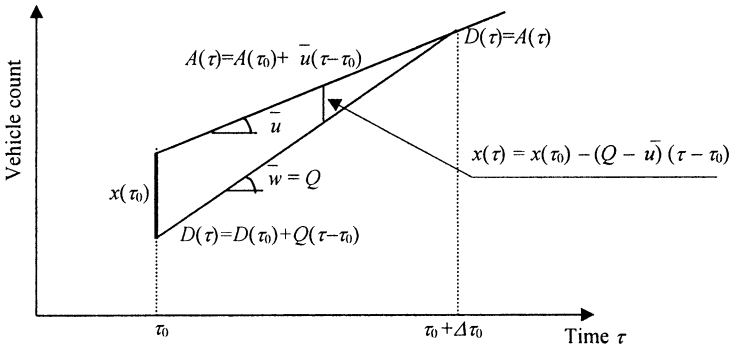


Fig. 2.A.10 Under-saturated queuing system.

After time $\tau_0 + \Delta\tau_0$ the queue length is zero and the exiting flow $\bar{w}$ is equal to the arrival flow $\bar{u}$:

$$x(\tau_0 + \Delta\tau_0) = 0 \tag{2.A.23}$$

$$\bar{w} = \bar{u}$$

$$D(\tau) = A(\tau) = A(\tau_0) + \bar{u}(\tau - \tau_0)$$

b) *Over-saturation*

When the arrival flow rate is larger than capacity, $\bar{u} \geq Q$, the system is *over-saturated.* In this case queue length linearly increases with time τ and the exiting flow is equal to the capacity (see Fig. 2.A.11):

$$x(\tau_0) = x(\tau_0) + (\bar{u} - Q)(\tau - \tau_0) \tag{2.A.24}$$

$$\bar{w} = Q$$

$$D(\tau) = D(\tau_0) + Q\,(\tau - \tau_0)$$

Comparing the eqns (2.A.22), (2.A.23) and (2.A.24) it is possible to formulate this general equation for calculating the queue length at generic time instant τ:

$$x(\tau) = \text{MAX}\ \{\ 0,\ [x(\tau_0) + (\ \bar{u} - Q\ )\,(\tau - \tau_0)]\ \} \qquad (2.A.25)$$

With the above results, any general case can be analyzed by modeling a sequence of periods during which arrival flow and capacity are constant. A relevant case is the analysis of delay at signalized intersections (periodical over-saturation conditions), as described in the section 2.A.3. (and also in the section 2.3.1.).

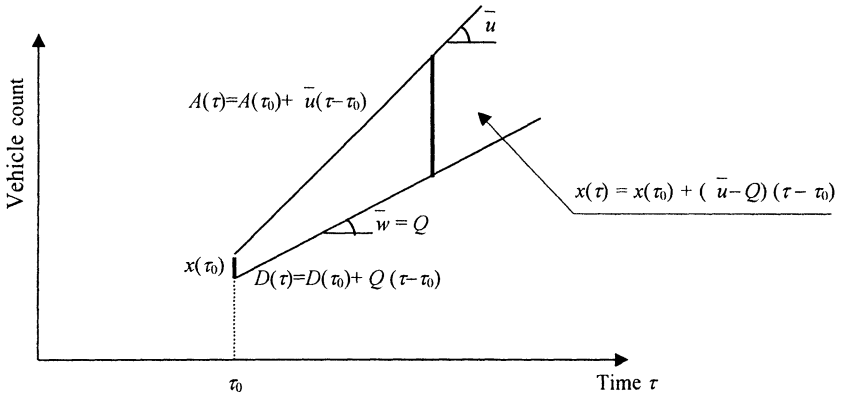


Fig. 2.A.11 Over-saturated queuing system.

The *delay* can be defined as the time needed for a user to leave the system (passing the server), accounting for the time spent in queue (pure waiting). Thus the delay is the sum of two terms:

$$tw = T_s + tw_q$$

where:

tw is the total delay;
T_s is the average service time (time spent at the server);
tw_q is the queuing delay (time spent in queue).

In under-saturated conditions ($\bar{u} < Q$) if the queue length at the beginning of period is zero (it remains equal to zero), the queuing delay is equal to zero, $tw_q(\ \bar{u})=0$, and the total delay is equal to the average service time:

$$tw(\ \bar{u}\ ) = T_s$$

In over-saturated conditions ($\bar{u} \geq Q$), the queue length, and respective delay, would tend to infinity in the theoretical case of stationary phenomenon lasting for an infinite time. In practice, however, over-saturated conditions last only for a finite period, T.

If the queue length is equal to zero at the beginning of the period, it will reach a value ($\bar{u} - Q$)$\cdot T$ at the end of period. Thus, the average queue over the whole period T is:

$$\bar{x} = \frac{(\bar{u} - Q)\, T}{2}$$

In this case the average queuing delay is $\bar{x} / Q$, and average total delay is:

$$tw(\bar{u}) = T_s + \frac{(\bar{u} - Q)\, T}{2Q} \tag{2.A.26}$$

The correspondent delay curve is reported in Fig. 2.A.12.

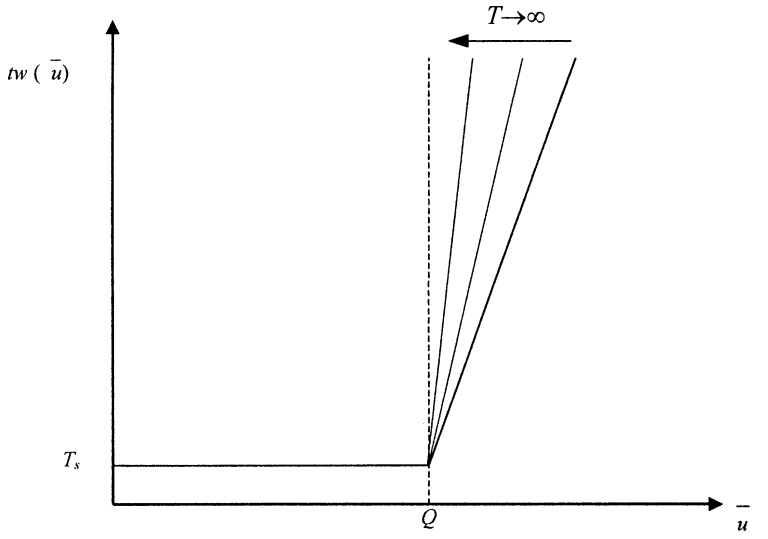


Fig. 2.A.12 Deterministic delay function at a server.

2.A.2.3. Stochastic models

Stochastic models arise when the variables of the problem (e.g. user arrivals, service times of server, etc.) cannot be assumed deterministic, as it is often the case, especially in traffic engineering.

If the system is *under-saturated*, it can be analyzed through (stochastic) queuing theory.

It is necessary to specify the following elements:

- the stochastic process describing the sequence of user arrivals (arrival pattern);
- the stochastic process describing the sequence of service times (service pattern);
- the queue discipline.

The characteristics of a queuing phenomenon can be redefined in the following concise notation:

$$a / b / c \ (d, e)$$

where:

a denotes the type of arrival pattern;
b denotes the type of service pattern;
c is the number of service channels;
d is the queue storage limit (∞, x_{max});
e denotes the queuing discipline.

Arrival and service processes are usually assumed to be stationary renewal processes, i.e. headways between successive arrivals and successive service times are independently distributed random variables with time-constant parameters. Let X be a random variable describing the queue length, and x the realization of X.

The symbol used for a and b positions refer to the random variables corresponding to the arrival and service times respectively, they may be:

D = deterministic;
M = negative exponential random variable;
E = Erlang random variable;
G = general distribution random variable.

The main queuing disciplines are:

FIFO = First In - First Out (i.e. service in order of arrival);
LIFO = Last In - First Out (i.e. the last user is the first served);
SIRO = Service In Random Order;
HIFO = High In - First Out (i.e. the user with the maximum value of an *indicator* is the first served.

In the following we will report the main results for the *M/M/*1 (∞, *FIFO*) and the *M/G/*1 (∞, *FIFO*) queue systems, which are commonly used for simulating transportation facilities, such as signalized intersections.

a) *M/M/1 (∞, FIFO) systems*
In this case the main parameters regulating the phenomenon are:

u the average arrival rate;
$Q = 1/T_s$ the service rate (or capacity) of the system;
u/Q the traffic intensity ratio or utilization factor.

In *under-saturated* conditions ($u/Q < 1$) the expected queue length can be calculated assuming that the arrivals are exponentially distributed; with this assumption the *expected value* and the *variance* of number of users in the system[7], X, can be obtained:

$$E[X] = \frac{\frac{u}{Q}}{1-\frac{u}{Q}} = \frac{u}{Q-u} \tag{2.A.27}$$

$$VAR[X] = \frac{\frac{u}{Q}}{\left(1-\frac{u}{Q}\right)^2} \tag{2.A.28}$$

The expected number of users in the system, $E[X]$, is the product of the average time in the system (expected value of delay), $E[tw]$, multiplied by arrival rate u. Then, $E[tw] = E[X]/u$ or:

$$E[tw] = \frac{1}{Q-u} \tag{2.A.29}$$

The expected time spent in queue, $E[tw_q]$, (or queuing delay) is given by the difference between the expected delay, $E[tw]$, and the service time $T_s = 1/Q$:

$$E[tw_q] = \frac{1}{Q-u} - \frac{1}{Q} = \frac{u}{Q(Q-u)} \tag{2.A.30}$$

Let X_q be the number of users in queue, then the expected queue length, $E[X_q]$, is the product of the expected queuing delay, $E[tw_q]$, multiplied by the arrival rate, u:

$$E[X_q] = \frac{u^2}{Q(Q-u)} \tag{2.A.31}$$

b) *M/G/1 (∞, FIFO) systems*
In this case the main results are the following:

$$E[X] = \frac{u}{Q}\left[1 + \frac{u}{2(Q-u)}\right]$$

$$E[tw] = \frac{1}{Q}\left[1 + \frac{u}{2(Q-u)}\right]$$

$$E[tw_q] = \frac{u}{2Q(Q-u)}$$

$$E[X_q] = \frac{u^2}{2Q(Q-u)}$$

2.A.3. Application to signalized intersections

Queuing and delay phenomena at signalized intersections can be obtained from queuing theory results reported in section 2.A.2. In fact, signalized intersections are a particular case of servers, for which the capacity is periodically equal to zero (when the signal is *red*). During such time the system is necessarily over-saturated. It is common to divide the cycle length into two time intervals (see also Fig. 2.3.8). The effective green time equals the green plus yellow time minus the lost time, during which departures occur at a constant service rate, given by the inverse of saturation flow. The effective red time is the difference between cycle length and the effective green time, during which no departures occur. Let

T_c be the cycle length for the whole intersection;
G be the effective green time for an approach;
$R = T_c - G$ be the effective red time for the approach;
$\mu = G/T_c$ is the effective green/cycle ratio for the approach.

The number of vehicles arriving at the approach during the time interval T_c are given by the following equation:

$$m_{IN}(\tau, \tau + T_c) = u \cdot T_c$$

The maximum number of the users that may leave the approach, during time interval T_c, is given by:

$$S \cdot G = S \cdot \mu \cdot T_c$$

where S is the saturation flow of the intersection approach (i.e. the capacity of the approach if the whole cycle were green).

Hence the actual capacity of the approach is given by:

$$Q = \frac{S \cdot G}{T_c} = \mu \cdot S$$

Thus, the approach can be defined *under-saturated* if:

$$\bar{u} \cdot T_c < S \cdot \mu \cdot T_c$$

that is:

$$\bar{u} < \mu S \qquad (2.A.32)$$

On the other hand the approach is defined *over-saturated* if:

$$\bar{u} \geq \mu S \qquad (2.A.33)$$

a) *Deterministic queuing models*

From equations (2.A.32) and (2.A.33) it is clear that the results discussed in section 2.A.2 hold for a queuing system representing a signalized intersection approach. Here Q is considered the green-time capacity ($Q = \mu S$) and the queue length given by (2.A.25) should be increased to include the vehicles that arrived during the effective red time R. This queue, also named *under-saturated queue*, shows a periodical trend over time, with a zero value at the beginning of effective red interval and at the end of the effective green time, and it assumes the maximum value at the end of the red interval (see Fig. 2.A.13).

The *under-saturated queue length*, $x_u(\tau)$, can be computed by equation (2.A.25) for a time period equal to R with zero capacity, and a period of time equal to G with capacity S (see Fig. 2.A.13). Then, $x_u(i \cdot T_c) = 0$ since the under-saturated queue length at the end of i-th cycle interval, $\tau = i \cdot T_c$, is equal to zero.

During an effective red time interval the under-saturated queue can be calculated by applying (2.A.25) with $Q = 0$, $\tau_0 = i \cdot T_c$, $x(\tau_0) = 0$:

$$x_u^R(\tau) = \bar{u}\,(\tau - i \cdot T_c) \qquad i \cdot T_c \leq \tau \leq i \cdot T_c + R \qquad (2.A.34)$$

The queue length reaches a maximum value at the end of the red-time, equal to:

$$x_u^R(i \cdot T_c + R) = \bar{u} R = \bar{u} (1 - \mu) T_c$$

During an effective green time interval, applying (2.A.25) with $Q = S$, $\tau_0 = i \cdot T_c + R$ and $x(\tau_0) = \bar{u} (1 - \mu) T_c$, the queue length is:

$$x_u^G(\tau) = \text{MAX} \{ 0, \bar{u} (1 - \mu) T_c - (S - \bar{u}) (\tau - i \cdot T_c - R) \} \qquad (2.A.35)$$

$$i \cdot T_c + R \le \tau \le i \cdot T_c + R + G$$

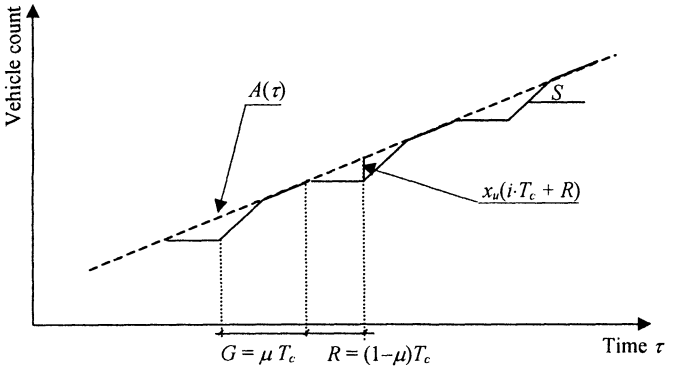


Fig. 2.A.13 Deterministic queue model for signalized intersections. Under-saturated condition.

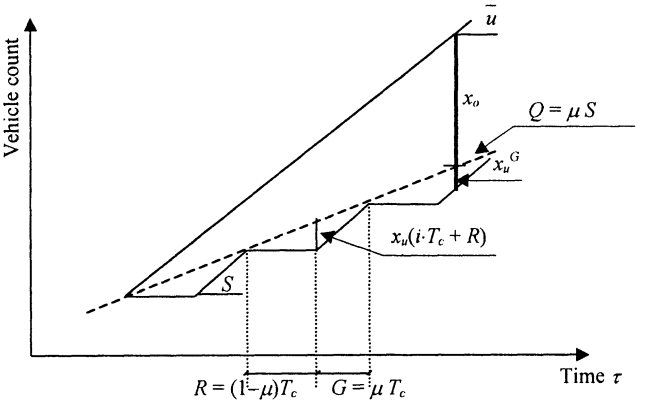


Fig. 2.A.14 Deterministic queue model for signalized intersections. Over-saturated condition.

The queue vanishes after a time $\Delta\tau_0$ given by (2.A.21):

$$\Delta\tau_0 = \frac{\bar{u}(1-\mu)T_c}{(\mu S - \bar{u})}$$

If over saturation occurs, $\bar{u} \geq Q = \mu S$, the *under-saturated queue length*, $x_u(\tau)$, is given by the equations (2.A.34) and (2.A.35) for an arrival flow equal to the capacity:

$$x_u^R(\tau) = \mu S(\tau - i \cdot T_c) \qquad i \cdot T_c \leq \tau \leq i \cdot T_c + R \tag{2.A.36}$$

$$x_u^G(\tau) = \mu S(1-\mu) T_c - S(1-\mu)(\tau - i \cdot T_c - R) \}$$
$$i \cdot T_c + R \leq \tau \leq i \cdot T_c + R + G \tag{2.A.37}$$

The *over-saturated queue length* can be computed with the queue obtained from (2.A.25) with $Q = \mu S$, $\tau_0 = 0$ and $x(\tau_0) = 0$ (see Fig. 2.A.14):

$$x_0(\tau) = (\bar{u} - \mu S)\, \tau$$

The *total queue length* is obtained by summing the over-saturated and the under-saturated queue lengths (see Fig. 2.A.13).

b) *Deterministic delay models*

Delays at signalized intersections can be studied separately for *under-saturated* and *over-saturated* conditions.

For under-saturated conditions, $\bar{u} < \mu S$, (the capacity of server is $Q = \mu S$) the average individual delay, tw_{US}, can easily be obtained from the evolution over time of the queue length, as described by equations (2.A.34) and (2.A.35):

$$tw_{US} = \frac{T_c[1-\mu]^2}{2[1-\bar{u}/S]} \tag{2.A.39}$$

In over-saturated conditions, $\bar{u} > \mu S$, as for the deterministic case, the queue length, and respective delay, would tend theoretically to infinity. In practice, however, over-saturation lasts only for a finite period of time, T, and the average delay, tw_{OS}, can be calculated from the evolution over time of queue length as described by equations (2.A.36), (2.A.37) and (2.A.38):

$$tw_{OS} = \frac{T_c[1-\mu]}{2} + \frac{T}{2}[(\bar{u}/\mu S) - 1] \tag{2.A.40}$$

Notice that the first term is the value of (2.A.39) for $\bar{u} = \mu S$. The delay for the arrival flows can be computed through equation (2.A.39) for $\bar{u} < \mu S$ and through (2.A.40) for $\bar{u} \geq \mu S$, as depicted in Fig. 2.A.15. Note that, unlike the Fig. 2.A.12, the diagram depicted in Fig. 2.A.15 shows an increase in the average delay also for flows below the capacity. This is due to the increase of the under-saturated delay expressed by (2.A.39).

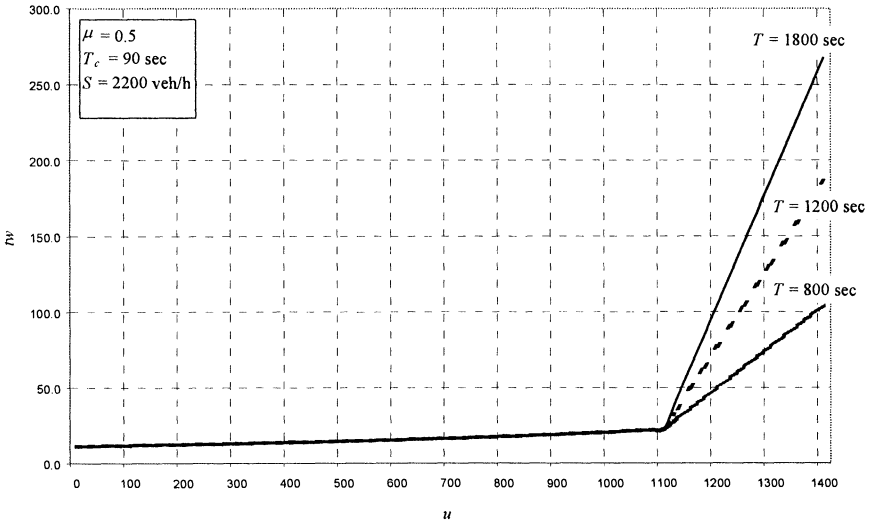


Fig. 2.A.15 Deterministic delay function at a signalized intersection.

c) *Stochastic delays models*

Stochastic delay models are based on the results of queuing theory. More precisely, a signalized intersection is considered to be a *M/G*/1 (∞, FIFO) system. Therefore, the average delay is (see section 2.A.2.3.):

$$tw^{st}{}_q(u) = \frac{(u/\mu S)^2}{2u(1-u/\mu S)} \tag{2.A.41}$$

d) *Total delay models*

The total delay equals the sum of the deterministic and the stochastic terms, and sometimes, terms calibrated through experimental observations.

Among the several models proposed in literature, the model proposed by Webster (see also section 2.3.1.2.) is commonly used for under-saturated condition.

$$tw(u) = \frac{T_c(1-\mu)^2}{2(1-u/S)} + \frac{(u/\mu S)^2}{2u(1-u/\mu S)} - 0.65(\mu S/u^2)^{1/3}(u/\mu S)^{2+\mu}$$

The first term is the deterministic delay, see eqn (2.A.39), the second is the delay due to random arrivals (2.A.41) and the third is a correction term based on numerical simulations. Delay tends to infinity for an arrival flow, u, approaching capacity μS (see Fig. 2.A.16). Thus, Webster's formula cannot be used to simulate delays for over-saturated signalized intersections. In order to calculate the delay in over-saturated conditions, different formulas have been proposed combing stationary models (e.g. Webster) with deterministic over-saturation models. The general form of these models is obtained by moving the asymptote from a vertical to an inclined position, as for over-saturation waiting time functions for toll barrier links (see Fig. 2.3.7). An example of such a model is the delay formula proposed by Akcelik, see eqn 2.4.12 and Fig. 2.3.10.

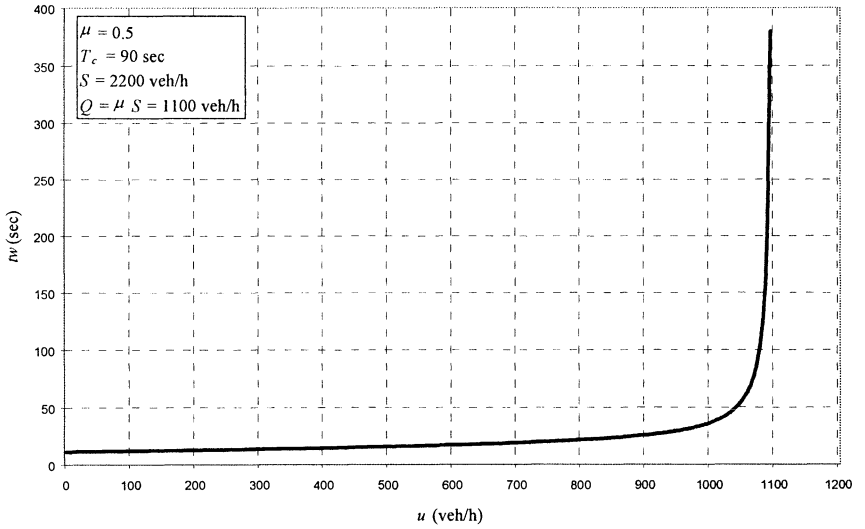


Fig. 2.A.16 Webster delay model.

Reference Notes

The application of network theory to the modeling of transportation supply systems can be found in most texts dealing with mathematical models of transportation systems, such Potts and Oliver (1972), Newell (1980), Sheffi (1985), Cascetta (1998), Ferrari (1996) and Ortuzar and Willumsen (1994). All of these, however, deal prevalently or exclusively with road networks. The presentation of general transportation supply model and its decomposition into sub models as described in Fig. 2.1.1 is original.

Performance models and the theory of road traffic flows are dealt with in several books and scientific papers. Among the former, Pignataro (1973), the ITE manual (1982), May (1990), Mc Shane and Roess (1990), the Highway Capacity Manual (1997), the relevant entries in the encyclopedia edited by Papageorgious (1991). Among the latter, the pioneering work of Webster (1958), later expanded in Webster and Cobbe (1966) and those of Catling (1977), Kimber, Marlow and Hollis (1977) Kimber and Hollis (1978), Robertson (1979), and Akcelik (1988) on waiting times at signalized intersections.

In the work of Drake, Shofer and May (1967) is reported a review of the main speed-flow-density relationships, and an example of their calibration. The linear model was proposed by Greenshields (1934). References of non-stationary traffic flow models are in part reported in the bibliographical note of Chapter 6. The second order model reported in the text is due to Payne (1971).

A review of the road network cost functions can be found in Branston (1976), Hurdle (1984) and Lupi (1996). The study of Cascetta and Nuzzolo (1982) contains experimental speed-flow relationship for urban roadways, reported in the text (equations 2.4.9). The cost function for parking links (equation 2.4.14) was proposed by Bifulco (1993).

Supply models for scheduled services have traditionally received less attention in the scientific community. The line representations of scheduled systems are described in Nguyen and Oakkittubi (1985), Ferrari (1996) and in Nuzzolo and Russo (1997).

Several authors, such as Seddon and Day (1974), Joliffe and Hutchinson (1975), Montella and Cascetta (1978) and Cascetta and Montella (1979) have studied the relationships between waiting times and service regularity in urban transit systems. Congested performance models discussed in section 2.3.2. have been proposed by Nuzzolo and Russo (1993), other models for the waiting time at congested bus stops are quoted in Bouzaiene-Ayari et al. (1998).

For a theoretical analysis of queuing theory, reference can be made to Newell (1971) and Kleinrock (1975).

Notes

(1) A distinction should be made between cost functions in micro-economics and in transportation systems theory. In the first case, the cost function is a relationship connecting the production cost of a good or service to the quantity produced and the costs of individual production factors. Cost functions in transportation systems relate the cost perceived by users in their trips. Transportation cost is therefore a cost of use rather than of production. The cost for producing transportation services is usually indicated as the service production cost, and similarly the functions relating it to the relevant quantities are called production cost functions.

(2) In general the accurate simulation of delays for coordinated networks of intersection is even more complex and it is typically accomplished through more detailed models.

(3) More in detail, it is assumed that the users' arrival is a Poisson process, i.e. the intervals between two successive arrivals are distributed according to a negative exponential variable.

(4) Expression (2.3.17) holds in principle when vehicle arrivals of all lines are completely irregular. In this case cumulated headways can still be modeled as a negative exponential random variable, with parameter equal to the inverse of the sum of line frequencies. In practice, however, expression (2.3.17) is often used also for intermediate values of θ.

(5) It is worth noting that the time mean speed is not less then the space mean speed, as it can be shown since the two speeds are related by the equation $\bar{v}_t = \bar{v}_s + \sigma^2 / \bar{v}_s$, where σ^2 is the variance of speed among vehicles. In Fig. 2.A.2 $\sigma^2 = 0$ hence $\bar{v}_t = \bar{v}_s$.

(6) Different vehicle types (motorcycle, personal car, truck, etc.) are reduced by an equivalence coefficient to a standard vehicle (for example personal car), see section 2.2.3.

(7) Note that the number of users in the system is different from the queue length (X_q) since it includes also the user being served.

3 RANDOM UTILITY THEORY

3.1. Introduction

In Chapter 1 it was stated that transport demand flows result from the aggregation of individual trips. Each trip is the result of several choices made by the users: travelers in passenger transportation or operators (manufacturers, shippers, carriers) in goods transport. Some traveler choices are made infrequently, such as where to reside and work and whether to own a vehicle or not. Other choices are made for each trip, these include whether to make a trip for a certain purpose at what time to what destination, with what mode, using what route. Each choice context, defined by available alternatives, evaluation factors and decision procedures, is usually known as a "choice dimension". Also, in most cases, choices concerning transport demand are made among a finite number of *discrete alternatives*.

Starting from these assumptions, many travel demand models described in the next chapter attempt to reproduce users' choice behavior [1] (behavioral models). The present chapter describes the mathematical models derived from random utility theory, which is the richest, and by far the most widely used [2] theoretical paradigm for the simulation of transport related choices and, more generally, choices among discrete alternatives. Within this paradigm, it is possible to specify several models, with various functional forms, applicable to a variety of contexts. It is also possible to study their mathematical properties and estimate their parameters using well established statistical methods.

It should be said that random utility models are not the only behavioral models that can be used to simulate transport related choices. Other models proposed in the literature are based on choice mechanisms, which violate one or more of the general hypotheses described in section 3.2. These models are usually referred to as "non compensatory" models since they do not allow the compensation of negative attributes with positive ones. Non-compensatory models are at present mostly research tools and are not widely used in practice. Furthermore, it has been shown that a properly specified random utility model can very often satisfactorily approximate the choice probabilities obtained with non-compensatory models.

In this chapter random utility models will be exemplified for personal mobility choices. The same models can be applied to simulate freight transport-related choices as will be seen in section 4.6. Section 3.2 introduces the general hypotheses underlying random utility models and section 3.3 describes their most widely used functional forms. Section 3.4 considers the problem of choice set modeling. Section

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

3.5 defines the Expected Maximum Perceived Utility variable and analyzes the mathematical properties of this variable and, of random utility models. Section 3.6 introduces the concept of elasticity of random utility models. Finally, section 3.7 analyses various aggregation procedures allowing the estimation of aggregate demand, starting from models simulating individual choices.

3.2. Basic assumptions

Random utility theory is based on the hypothesis that every individual is a *rational decision-maker*, maximizing utility relative to his/her choices. Specifically, the theory is based on the following assumptions:

a) the generic decision-maker i, in making a choice, considers m_i mutually exclusive alternatives which make up his/her *choice set* I^i. The choice set may be different for different decision-makers (for example, in the choice of transport mode, the choice set of an individual without driving license and/or car obviously does not include the alternative "car as a driver");
b) decision-maker i assigns to each alternative j from his/her choice set a perceived utility, or "attractiveness" U^i_j and selects the alternative maximizing this utility;
c) the utility assigned to each choice alternative depends on a number of measurable characteristics, or *attributes*, of the alternative itself and of the decision-maker, $U^i_j = U^i(\boldsymbol{X}^i_j)$, where $\boldsymbol{X}^i_j$ is the vector of the attributes relative to alternative j and to decision-maker i;
d) the utility assigned by decision-maker i to alternative j is not known with certainty by an external observer (analyst), because of a number of factors that will be described later and must therefore be represented by a random variable.

On the basis of the above assumptions, it is not usually possible to predict with certainty the alternative that the generic decision-maker will select. However, it is possible to express the probability of selecting alternative j conditional on his/her choice set I^i, as the probability that the perceived utility of alternative j is greater than that of all the other available alternatives:

$$p^i[j/I^i] = Pr[U^i_j > U^i_k \quad \forall k \neq j, k \in I^i] \tag{3.2.1}$$

The perceived utility U^i_j can be expressed by the sum of the *systematic utility* V^i_j, which represents the mean or the expected value of the utilities perceived by all decision-makers having the same choice context as decision-maker i (same alternatives and attributes), and a *random residual* ε^i_j, which is the (unknown) deviation of the utility perceived by the user i from this value:

$$U^i_j = V^i_j + \varepsilon^i_j \quad \forall j \in I^i \tag{3.2.2a}$$

with:

$$V_j^i = E[U_j^i] \quad \sigma_{i,j}^2 = Var[U_j^i]$$

and therefore:

$$E[V_j^i] = V_j^i \qquad Var[V_j^i] = 0$$

$$E[\varepsilon_j^i] = 0 \qquad Var[\varepsilon_j^i] = \sigma_{i,j}^2$$

Replacing the expression (3.2.2a) in (3.2.1) we have:

$$p^i[j / I^i] = Pr[V_j^i - V_k^i > \varepsilon_k^i - \varepsilon_j^i \quad \forall k \neq j, k \in I^i] \tag{3.2.3a}$$

From (3.2.3a) it follows that the choice probability of an alternative depends on the systematic utilities of all competing (available) alternatives, and on the joint probability law of random residuals ε_j.

Random utility models and relative variables can be represented by introducing a vector notation. Let

$\boldsymbol{p}^i$ be the vector of choice probabilities, of dimension $(m_i \times 1)$, with elements $p^i[j]$;
$\boldsymbol{U}^i$ be the vector of perceived utilities of dimension $(m_i \times 1)$, with elements U_j^i;
$\boldsymbol{V}^i$ be the vector of systematic utilities values of dimension $(m_i \times 1)$, with elements V_j^i;
$\boldsymbol{\varepsilon}^i$ be the vector of random residuals, of dimension $(m_i \times 1)$, with elements ε_j^i;
$f(\boldsymbol{\varepsilon})$ be the joint probability density function of random residuals;
$F(\boldsymbol{\varepsilon})$ be the joint probability distribution function of random residuals.

Expression (3.2.2a) can therefore be written in vector notation as:

$$\boldsymbol{U}^i = \boldsymbol{V}^i + \boldsymbol{\varepsilon}^i \tag{3.2.2b}$$

In general, the choice model (3.2.3a) can be seen as a function, known as a *choice function*, associating a vector of choice probabilities to each vector $\boldsymbol{V}^i$ of systematic utilities for a given probability law of random residuals:

$$\boldsymbol{p}^i = \boldsymbol{p}^i(\boldsymbol{V}^i) \quad \forall \boldsymbol{V}^i \in E^{m_i} \tag{3.2.3b}$$

A random utility model is said to be *additive* if the joint probability density function of random residuals, $f(\boldsymbol{\varepsilon})$, or its parameters, is not dependent on the vector $\boldsymbol{V}$ of systematic utilities:

$$f(\boldsymbol{\varepsilon}/\boldsymbol{V}) = f(\boldsymbol{\varepsilon}) \quad \forall \boldsymbol{\varepsilon} \in E^{m_i}$$

It follows immediately from expression (3.2.3a) that for additive models the choice probabilities of each alternative do not vary if a constant V_0 is added to the systematic utility of all the alternatives:

$$p^i[j/I^i]=Pr[V^i_j+V_0-V^i_k-V_0>\varepsilon^i_k-\varepsilon^i_j]=Pr[V^i_j-V^i_k>\varepsilon^i_k-\varepsilon^i_j]\ \forall\, k\neq j;\ j,\ k\in I^i \quad (3.2.4)$$

From the previous expression it also results that, in the case of additive models, choice probabilities depend on the differences between systematic utilities, known as the relative systematic utilities $V_j - V_h$, relative to any reference alternative h.

Before describing some of the random utility models derived from various assumptions on the joint probability functions of random residuals, some further general remarks on the implications of the hypotheses introduced so far should be made.

The variance-covariance matrix of random residuals. In general, a variance-covariance matrix Σ is symmetric and positive semidefinite. When the variance of each random residual, ε_k, is null, $\sigma_{kk} = 0$, all the covariances must be null, $\sigma_{kh} = 0$ $\forall h$, therefore the variance-covariance matrix is null, $\Sigma = 0$; in this case we obtain the *deterministic* choice model whose properties are described in section 3.5. If the variance-covariance matrix is not null, $\Sigma \neq 0$, a *non-deterministic* choice model is obtained. In this case, it is usually assumed that the variance $\sigma_{kk} = \sigma_k^2$ of each random residual, ε_k, is strictly positive, $\sigma_{kk} > 0$, and that the random residuals are imperfectly correlated $(\sigma_{kh})^2 < \sigma_k^2 \sigma_h^2$; i.e. the rows (or columns) of Σ are pairwise linearly independent. These conditions are equivalent to assuming that the variance-covariance matrix is not singular, $|\Sigma| \neq 0$ in addition to being non null, $\Sigma \neq 0$. In this case the models are defined as *probabilistic*[3], and the choice function $p=p(V)$ is continuous with continuous first partial derivatives.

The set of available alternatives I^i, or choice set, influences significantly the choice probabilities, as can be seen from equations (3.2.1) and (3.2.3a). If the choice set I^i of the single decision-maker is known, the definition of choice probability (3.2.1) can be applied directly. However, it often happens that the analyst does not know exactly the generic decision-maker's choice set. In this case the problem can be handled with different levels of approximation as will be seen in section 3.4.

Expression of systematic utility. Systematic utility is the mean of the perceived utility among all individuals who have the same attributes; it is expressed as a function $V^i_j(X^i_{kj})$ of attributes X^i_{kj} relative to the alternatives and the decision-maker. Although the function $V^i_j(\boldsymbol{X}^i_j)$ may be of any type, for analytical and statistical convenience, it is usually assumed that the systematic utility V^i_j is a linear function in the coefficients β_k of the attributes X^i_{kj} or of their functional transformations $f_k(X^i_{kj})$:

$$V^i_j(\boldsymbol{X}^i_j) = \sum_k \beta_k X^i_{kj} = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}^i_j \quad (3.2.5a)$$

or

$$V^i_j(\boldsymbol{X}^i_j) = \sum_k \beta_k f_k(X^i_{kj}) = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{f}(\boldsymbol{X}^i_j) \quad (3.2.5b)$$

One useful parametric functional transformation for non-negative variables is the Box Cox one:

$$x_k \rightarrow (x_k^{\lambda_k} - 1)/\lambda_k \qquad \text{if } \lambda_k \neq 0$$
$$x_k \rightarrow \log(x_k) \qquad \text{if } \lambda_k = 0$$

where λ_k is an unknown parameter. This transformation defines a family of functions that includes, as special cases, the linear (λ_k=1), the power (λ_k≠0) and the logarithmic (λ_k=0) transformations. The numerical coefficients β_k in expressions (3.2.5.a) and (3.2.5.b) can be estimated using various statistical techniques described in Chapter 8. The Box-Cox transformation introduces some difficulties in the estimation process due to the non linearity of the utility function in the λ that can be avoided by iteratively estimate the model for different fixed value of the λ.

The attributes contained in the vector X_j^t can be classified in different ways. The attributes related to the service offered by the transport system are known as *level of service* or *performance attributes* (times, costs, service frequency, comfort, etc.). Attributes related to the land-use of the study area (for example, the number of shops or schools in each zone) are known as *activity system attributes*. Attributes related to the decision-maker or his/her household (income, holding a driving license, number of cars in the household, etc.) are usually referred to as *socio-economic attributes*.

Attributes of any type might be *generic*, if they are included in the systematic utility of more than one alternative in the same form and with the same coefficient β_k. They are *specific*, if included with different functional forms and/or coefficients in the systematic utilities of different alternatives. A dummy variable is usually introduced into the systematic utility of the generic alternative j; its value is one for alternative j and zero for the others. This variable is usually denoted *Alternative Specific Attribute* (ASA) or "modal preference" attribute[4], and its coefficient β is known as the Alternative Specific Constant (ASC). The ASA is a kind of "constant term" in the systematic utility which can be seen as the difference between the mean utility of an alternative and that explained by the other attributes X_{kj}.

From expression (3.2.4) it results that the choice probabilities of additive models depend on the differences of the ASC of each alternative j with respect to a reference alternative h. If the Alternative Specific Constants should appear in the systematic utilities of all alternatives, there would be infinite combinations of such constants which would result in the same values of the choice probabilities. For this reason, in order to avoid problems in the estimation of coefficients β, in the specification of additive models, ASA's are introduced at most into the systematic utilities of all the alternatives except one.

An elementary example of systematic utilities related to transportation mode choice is given in Fig. 3.2.1. Many other examples will be given in the following chapters.

$V_{walking} = \beta_1 t_{wl}$
$V_{auto} = \beta_1 t_{wla} + \beta_2 t_{ba} + \beta_3 mc_a + \beta_4 AVAIL + \beta_5 INC + \beta_6 AUTO$
$V_{bus} = \beta_1 t_{wlb} + \beta_2 t_{bb} + \beta_3 mc_b + \beta_7 t_{wb} + \beta_8 BUS$

ALTERNATIVE SPECIFIC ATTRIBUTES (ASA)	LEVEL OF SERVICE ATTRIBUTES		SOCIO-ECONOMIC ATTRIBUTES	
AUTO *BUS*	t_b t_w t_{wl} mc	= time on board (generic) = waiting time at stop (specific) = walking time (generic) = monetary cost (generic)	*AVAIL* *INC*	= n°auto/n°licenses = disposable income of the household

Fig. 3.2.1 Specification of systematic utilities and classification of attributes.

The utility of an alternative can be considered dimensionless, or expressed in arbitrary measurement units (*util*). From expression (3.2.5) it results that, in order to sum attributes expressed in various units (for example, times and costs) the relative coefficients β_k have to be expressed in measurement units inverse to those of the attributes themselves (for example time^{-1} and cost^{-1}). Coefficients β are sometimes denoted as reciprocal substitution coefficients since they allow to evaluate the reciprocal "exchange rates" between attributes. This point will be expanded in Chapter 4.

Randomness of perceived utilities. The difference between the perceived utility for a decision-maker and the systematic utility common to all decision-makers with equal values of the attributes, can be attributed to several factors related both to the model (a,b,c) and to the decision-maker (d,e). These are:

a) measurement errors of the attributes in the systematic utility. Level-of-service attributes are often computed through a network model and are therefore subject to modeling and aggregation (zoning) errors; some attributes are intrinsically variable and their average value is considered;
b) omitted attributes that are not directly observable, difficult to evaluate or not included in the attribute vector (e.g., travel comfort or the reliability of total travel time);
c) presence of instrumental attributes that replace the attributes actually influencing the perceived utility of alternatives (e.g., modal preference attributes replacing the variables of comfort, privacy, image, etc. of a certain transport mode; the number of commercial operators operating in a given zone replacing the number and variety of shops);
d) dispersion among decision-makers, or variations in tastes and preferences among decision-makers and, for the individual decision-maker, over time. Different decision-makers with equal attributes might have different utility values or different values of the reciprocal substitution coefficients β_k according to personal preferences (e.g. walking distance is more or less disagreeable to different people). The same decision-maker might weigh an attribute differently

in different decision contexts (e.g. according to different physical or psychological conditions);

e) errors in the evaluation of attributes by the decision-maker (e.g. erroneous estimation of travel time).

From the above discussion, it results that the more accurate the model (the more attributes included in the systematic utilities, the more precise their calculation, etc.) the lower should be the variance of random residuals ε_j. Experimental evidence confirms this conjecture.

3.3. Some random utility models

Various specifications of random utility models can be derived from the general hypotheses presented in the previous section by assuming different joint probability distribution functions for the random residuals ε_j[5] in expression (3.2.3a). This section describes the random utility models that are most widely used in travel demand modeling. Models are introduced in order of increasing generality and analytical complexity. Section 3.3.1 will describe the Multinomial Logit (or MNL) model, which is the simplest functional form. Subsequently, progressive generalizations of the MNL to the Single-Level Hierarchical or Nested Logit model (section 3.3.2), to the Multi-Level Hierarchical or Tree Logit model (section 3.3.3), to the Cross Nested Logit model (section 3.3.4), and to the Generalized Extreme Value (GEV) model (section 3.3.5) are described. Each of these models includes the MNL as a special case and can be obtained from the GEV model. Finally, section 3.3.6 describes the Probit model and section 3.3.7 introduces the Hybrid Logit Probit model.

3.3.1. The Multinomial Logit model

The Multinomial Logit (MNL) model is the simplest random utility model. It is based on the assumption that the random residuals ε_j are independently and identically distributed (i.i.d.) according to a Gumbel random variable (r.v.) of zero mean and parameter θ[6]. The marginal probability distribution function of each random residual is given by:

$$F_{\varepsilon_j}(x) = Pr[\varepsilon_j \le x] = exp[-exp(-x/\theta - \Phi)] \tag{3.3.1}$$

where Φ is the Eulero constant ($\Phi \approx 0.577$). In particular, mean and variance of the Gumbel variable expressed by (3.3.1) are respectively:

$$\begin{aligned} &E[\varepsilon_j] = 0 && \forall j \\ &Var[\varepsilon_j] = \sigma_\varepsilon^2 = \frac{\pi^2\theta^2}{6} && \forall j \end{aligned} \tag{3.3.2}$$

Further characteristics of the Gumbel r.v. are given in Appendix 3.B.

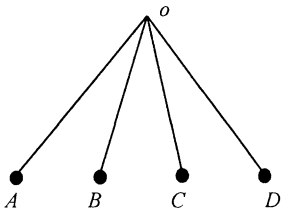
Furthermore the independence of the random residuals implies that the covariance between any pair of residuals is null:

$$Cov[\varepsilon_j, \varepsilon_h] = 0 \qquad \forall j, h \in I \tag{3.3.3}$$

From this it can be deduced that the perceived utility U_j, sum of a constant V_j and of the r.v. ε_j, is also a Gumbel random variable with probability distribution function, mean and variance given by:

$$F_{Uj}(x) = Pr[U_j \le x] = Pr[\varepsilon_j \le x - V_j] = exp[-exp(-(x - V_j)/\theta - \Phi)]$$
$$E[U_j] = V_j \qquad Var[U_j] = \frac{\pi^2\theta^2}{6} \tag{3.3.4}$$

On the basis of the hypotheses on the residuals ε_j, and therefore on the perceived utilities U_j, the residuals variance-covariance matrix, Σ_ε, for the available m alternatives, is a diagonal matrix proportional by σ_ε^2 to the identity matrix. Fig. 3.3.1 shows a graphic representation of the assumptions made on the distribution of random residuals in the Multinomial Logit model and the Variance-Covariance matrix in the case of four choice alternatives. This representation, known as choice tree, should be compared to that of the Hierarchical Logit models described in the following sections.



$$\sum_\varepsilon = \sigma_\varepsilon^2 \boldsymbol{I} = \frac{\pi^2\theta^2}{6} \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Fig. 3.3.1 Choice tree and variance-covariance matrix of a Multinomial Logit model.

The Gumbel variable has an important property known as *stability with respect to maximization*. The maximum of independent Gumbel variables of equal parameter θ is also a Gumbel variable of parameter θ. In other words, if U_j are independent Gumbel variables of equal parameter θ but with different means V_j, the variable U_M:

$$U_M = max_j \{U_j\}$$

is again a Gumbel variable with parameter θ and mean V_M given by:

$$V_M = E[U_M] = \theta ln \sum_j exp(V_j / \theta) \tag{3.3.5}$$

The variable V_M is denominated *Expected Maximum Perceived Utility* (*EMPU*)[7] or *inclusive utility* and the variable Y to this proportional, because of its analytical structure, is denominated "*logsum*":

$$Y = ln \sum_j exp(V_j / \theta)$$

Stability with respect to maximization makes the Gumbel variable a particularly convenient assumption for the distribution of residuals in random utility models. In fact, under the assumptions made, the probability of choosing alternative j among those available (1,2, ...,m) $\in I$, given by (3.2.4), can be expressed[8] in closed form as:

$$p[j] = \frac{exp(V_j / \theta)}{\sum_{i=1}^{m} exp(V_i / \theta)} \tag{3.3.6}$$

Expression (3.3.6) defines the *Multinomial Logit* model, which is the simplest and one of the most widely used random utility models. Under the common assumption that the parameter θ is independent of the systematic utility, the MNL model is additive (see section 3.5) and has certain important properties that will be described in the following.

Dependence on the differences among systematic utilities[9]. In the case of only two alternatives (A and B), the MNL model (3.3.6) is called Binomial Logit and can be expressed as:

$$p[A] = \frac{exp(V_A / \theta)}{exp(V_A / \theta) + exp(V_B / \theta)} = \frac{1}{1 + exp[(V_B - V_A / \theta)]}$$

The choice probability of alternative A depends on the difference between the systematic utilities. Furthermore, as shown in Fig. 3.3.2, this choice probability is equal to 0.5 if the two alternatives have equal systematic utilities (V_B-V_A=0). It has an S-shaped emi-symmetric diagram for positive and negative values of V_B-V_A. In addition, it tends to one as V_B-V_A tends to -∞ (alternative A has a systematic utility infinitely greater than that of B) while it tends to zero as V_B-V_A tends to +∞. The rate of variation of the choice probability of A with respect to variations of V_B-V_A, is larger for values of V_B-V_A close to zero where it is almost linear, and is the larger the smaller is the variance of random residuals (parameter θ). As the absolute value of V_B-V_A increases, $p[A]$ shows a flex and becomes sub-horizontal; for large differences V_B-V_A the variations of choice probability have low sensitivity to the variations of V_B-V_A.

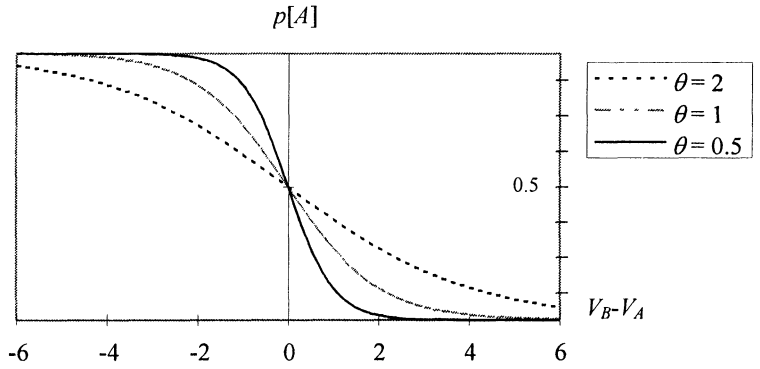


Fig. 3.3.2 Diagram of choice probability $p[A]$ of a Binomial Logit model.

Similar considerations apply to the more general case of the Multinomial Logit model. From expression (3.3.6) and from the general results on additive random utility models, it results that the choice probability of any alternative depends on the differences between the systematic utilities of all other alternatives. In fact, by dividing numerator and denominator of (3.3.6) by $\exp(V_j/\theta)$ it results:

$$p[j] = \frac{1}{1 + \sum_{h \neq j} exp[(V_h - V_j)/\theta]}$$

It follows that if Alternative Specific Attributes (ASA) and Coefficients (ASC) are introduced in each of the m alternatives, choice probabilities are equal to those that would be obtained if the ASA were introduced in all the alternatives except one and the relative coefficients were replaced with the differences with respect to the eliminated ASA coefficient. If the ASC of alternative j is denoted with β_j, and the remaining part of the systematic utility with V'_j, the Multinomial Logit model becomes:

$$p[j] = \frac{exp[(\beta_j + V'_j)/\theta]}{\sum_{h=1}^{m} exp[(\beta_h + V'_h)/\theta]} = \frac{exp(V'_j/\theta)}{exp(V'_j/\theta) + \sum_{h \neq j} exp[((\beta_h - \beta_j) + V'_h)/\theta]}$$

Influence of residual variance. From equation (3.3.6) it follows that a smaller variance of random residuals and a smaller parameter θ correspond to a larger choice probability for the maximum systematic utility alternative. This probability tends to one (deterministic utility model) as the variance tends to zero. On the other hand, as the variance of residuals increases, exponents V_j/θ tend to the same value (zero) and the different alternatives tend to have the same choice probability equal to

$1/m$. The effect of random residuals variance, is illustrated in Fig. 3.3.2 and numerically in Fig. 3.3.3 for two choice alternatives corresponding to two routes with attributes given by travel time (t) and monetary cost (mc).

$$p[A] = \frac{exp[(-0{,}1 \cdot t_A - 1 \cdot mc_A)/\theta]}{exp[(-0{,}1 \cdot t_A - 1 \cdot mc_A)/\theta] + exp[(-0{,}1 \cdot t_B - 1 \cdot mc_B)/\theta]}$$

$t_A = 20$ min $\quad c_A = 3.6$ unit $\quad V_A = -5.6$
$t_B = 40$ min $\quad c_B = 0.6$ unit $\quad V_B = -4.6$

	$\theta = 10$	$\theta = 1$	$\theta = 0{,}5$
p_A	0,48	0,27	0,12
p_B	0,52	0,73	0,88

Fig. 3.3.3 Effect of the variance of random residuals on choice probabilities for a Binomial Logit model.

Independence from Irrelevant Alternatives. From expression (3.3.6) another general property of the Logit model can easily be deduced. Choice probability ratios between any two alternatives are constant and independent of the number and systematic utility of other choice alternatives:

$$p[j]/p[h] = exp(V_j/\theta)/exp(V_h/\theta) \qquad (3.3.7)$$

This property known in the literature as Independence from Irrelevant Alternatives (IIA) can sometimes lead to unrealistic results. Consider, for example, the case of choice between two alternatives A and B of equal systematic utility. In this case the probability of choosing each alternative calculated with the Logit model (3.3.6) is 0.50 and the ratio between the probability of choosing A and B is equal to one:

$$p[A]/p[B] = exp(V_A/\theta)/exp(V_B/\theta) = 1$$

Suppose that a third alternative C is added to the choice set. Alternative C has an equal systematic utility but is very similar to alternative B. Imagine the choice between transport modes where alternative A is the car and alternative B is a bus. A notional third alternative C is introduced consisting in a new bus line which runs to the same timetable and makes the same stops as B. In this case, the ratio between the probability of choosing car A, and bus B, because of the IIA property, remains equal to one. Therefore each of the three alternatives would have a probability of being chosen of 1/3. Thus, the probability of choosing the car would change from 0.50 to 0.33 due to the fictitious increase in choice alternatives. This result is clearly paradoxical and derives, in the case described, from the lack of realism of the basic assumptions of the Logit model; namely that alternatives are distinctly perceived by the decision-maker and that their random residuals are independent. A more realistic choice model can be obtained by introducing a covariance between the random residuals of alternatives B and C, as will be seen in the following sections. In

general, in a Multinomial Logit model, any change in the characteristics of a given alternative is such that the variation of choice probabilities of this alternative implies proportional variations of the choice probabilities of all other alternatives.

In applications, the Multinomial Logit model should be used with sufficiently distinct choice alternatives for which the assumption of independent random residuals is plausible.

3.3.2. The Single-Level Hierarchical Logit model

The Hierarchical Logit model[10] allows to overcome partially the assumption of independent random residuals underlying the Multinomial Logit model. At the same time it retains a closed analytical expression.

For simplicity of exposition, this section deals with the simpler case of a single level of hierarchy, with equal parameters. The model is also introduced referring to a graphic representation of the choice process and a decomposition scheme of random residuals. These assumptions are not necessary and will be relaxed in the next section dealing with general Hierarchical Logit models and in section 3.3.5 dealing with Generalized Extreme Value models.

Suppose that the decision-maker's choice set I is subdivided into non-overlapping subsets $I_1, I_2,..., I_k,...$ so that the utility function of the generic alternative j, belonging to the subset I_k, can be expressed[11] as:

$$U_j = V_j + \varepsilon_j = V_j + \eta_k + \tau_{j/k} \quad \forall j \in I_k \quad \forall k \tag{3.3.8}$$

with

$$E[\varepsilon_j] = E[\eta_k] = E[\tau_{j/k}] = 0$$

$$Cov[\eta_k, \eta_h] = Cov[\eta_k, \tau_{j/k}] = Cov[\tau_{j/k}, \tau_{i/k}] = 0$$

It is assumed that the global random residual ε_j can be decomposed into the sum of two random variables of zero mean. The first, η_k takes on the same value for all the alternatives belonging to the same group, though it can assume different values for different groups. The second, $\tau_{j/k}$, takes on different values for each alternative. Also, it is assumed that the variables η_k and $\tau_{j/k}$ are statistically independent. These assumptions imply that the decision-maker perceives the alternatives belonging to the same group as similar; this similarity is modeled introducing a covariance among the random residuals of these alternatives. In mode choice example, the available modes can be divided into two groups: public modes (bus and train) and private modes (car and motorbike). The assumption (3.3.8) implies that the decision-maker perceives the modes belonging to the same group as being similar since they share a number of attributes (flexibility, privacy, etc.).

The utility structure and the choice mechanism corresponding to a Single-Level Hierarchical Logit model can be represented by a choice tree shown in Fig. 3.3.4. On the choice tree, "elementary" choice alternatives (e.g transport modes) correspond to nodes with no exit links ("leaves" of the tree), the root node "o" has

no entering links. The intermediate nodes k, one for each group, represent compound alternatives or groups of elementary alternatives. The random residuals η_k and $\tau_{j/k}$ are associated to the branches corresponding to groups and to single alternatives respectively.

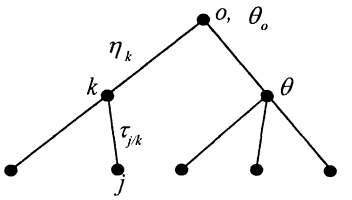


Fig. 3.3.4 Choice tree of a Single-Level Hierarchical Logit model.

The choice tree can be seen as the representation of a hypothetical choice process. The decision-maker, starting from the root node, first chooses group k from those available (represented by nodes linked to the root) and then the elementary alternative j from those belonging to group k (represented by the leaves connected to the node k). The expression of the overall choice probability of the generic alternative $p[j]$ is obtained as the product of the probability $p[j/k]$ of choosing elementary alternative j within group k (lower level), multiplied by the probability $p[k]$ of choosing group k (upper level). The name of the model is derived, in fact, from this probability structure:

$$p[j] = p[j/k] \cdot p[k] \tag{3.3.9}$$

To specify the probabilities in (3.3.9) further assumptions on the distribution of random residuals must be introduced. For the Single-Level Hierarchical Logit model it is assumed that the random residuals relative to the alternatives available at each decision node are identically and independently distributed (i.i.d.) Gumbel random variables. More precisely, residuals $\tau_{j/k}$ are i.i.d. Gumbel variables with zero mean and parameter θ for all groups k and all alternatives j. The perceived utility associated with alternative j in the choice among those belonging to group k, $U_{j/k}$, can be expressed as:

$$\begin{aligned} &U_{j/k} = V_j + \tau_{j/k} && \forall j \in I_k,\ \forall k \\ &E[\tau_{j/k}] = 0 && \forall j \in I_k,\ \forall k \\ &Var[\tau_{j/k}] = \pi^2\theta^2 / 6 && \forall j \in I_k,\ \forall k \end{aligned} \tag{3.3.10}$$

Under these assumptions the conditional choice probability of the elementary alternative j can be expressed as:

$$p[j/k] = Pr[U_{j/k} > U_{i/k}] = \\ = Pr[V_j - V_i > \tau_{i/k} - \tau_{j/k}] \qquad \forall i \in I_k, i \neq j \tag{3.3.11}$$

and given the assumptions on the distribution of residuals $\tau_{j/k}$, probability (3.3.11) results in a Multinomial Logit model:

$$p[j/k] = \frac{exp(V_j/\theta)}{\sum_{i \in I_k} exp(V_i/\theta)} \tag{3.3.12}$$

At the upper level, the choice is made among groups of alternatives, with each group k being considered as a compound alternative. The probability $p[k]$ is equivalent to the probability of choosing an elementary alternative belonging to group k. This probability is obtained by assigning to group k an inclusive perceived utility U^*_k equal to the utility of the most attractive alternative, i.e. the maximum utility of all the elementary alternatives belonging to the group:

$$U^*_k = max_{j \in I_k}\{U_j\} = max_{j \in I_k}\{V_j + \tau_{j/k}\} + \eta_k \tag{3.3.13}$$

As stated earlier, the maximum of independently distributed Gumbel variables with the same parameter θ is also distributed as a Gumbel variable of parameter θ and mean equal to:

$$V^*_k = E[U^*_k] = E[max_{j \in I_k}\{V_j + \tau_{j/k}\}] = \theta ln \sum_{j \in I_k} exp(V_j/\theta) = \theta Y_k \tag{3.3.14}$$

where V^*_k is the Expected Maximum Perceived Utility (EMPU) or inclusive utility and Y_k is the logsum variable. In the expression of the perceived utility (3.3.13) the r.v. $max(V_j + \tau_{j/k})$ can be replaced by its expected value plus the deviation from this value which is a another zero mean Gumbel variable τ^*_k[12], of parameter θ:

$$U^*_k = \theta Y_k + \tau^*_k + \eta_k = \theta Y_k + \varepsilon^*_k \tag{3.3.15}$$

Thus, the perceived utility of group k has a mean value θY_k and a deviation ε^*_k, which is the sum of the two zero mean random variables τ^*_k and η_k.

The basic assumption of the Hierarchical Logit model is that at each choice level the random residuals of the available alternatives are i.i.d. Gumbel variables; i.e. it is assumed that the ε^*_k are i.i.d. Gumbel variables of zero mean and parameter θ_o:

$$E[\varepsilon^*_k] = 0 \qquad \forall k \\ Var[\varepsilon^*_k] = \pi^2 \theta_o^2 / 6 \qquad \forall k \tag{3.3.16}$$

In accordance with this assumption, the choice probability of group k is expressed with a Multinomial Logit model. In fact:

$$p[k] = Pr[U_k^* > U_h^*] = Pr[\theta Y_k - \theta Y_h > \varepsilon_h^* - \varepsilon_k^*] \quad \forall h \neq k$$

and, given the results of the previous section:

$$p[k] = \frac{exp(\theta Y_k / \theta_o)}{\sum_h exp(\theta Y_h / \theta_o)} = \frac{exp(\delta Y_k)}{\sum_h exp(\delta Y_h)} \tag{3.3.17}$$

where δ is the ratio of parameters θ and θ_o associated to the two choice levels:

$$\delta = \theta / \theta_o \tag{3.3.18}$$

Replacing expressions (3.3.12) and (3.3.17) in (3.3.9), the choice probability of the generic alternative j can be obtained:

$$p[j] = p[j/k] \cdot p[k] = \frac{exp(V_j / \theta)}{\sum_{i \in I_k} exp(V_i / \theta)} \cdot \frac{exp(\delta Y_k)}{\sum_h exp(\delta Y_h)} \tag{3.3.19}$$

Variances and covariances of the random residuals ε_j of the overall perceived utility (3.3.8) can also be derived. The variance of ε_j coincides with that of the random residual ε^*_k since the two variables are the sum of the same variable (η_k) and of another independent Gumbel variable (τ^*_k and $\tau_{j/k}$ respectively) with zero mean and the same parameter θ. Therefore:

$$Var[\varepsilon_j] = Var[\varepsilon_k^*] = \pi^2 \theta_o^2 / 6 \qquad \forall j \tag{3.3.20}$$

The variance of random residuals ε_j is constant for all alternatives. There is also a positive covariance between the random residuals of any pair of alternatives belonging to the same group. In fact:

$$\begin{aligned} Cov[\varepsilon_i, \varepsilon_j] &= E[(\eta_k + \tau_{i/k}) \cdot (\eta_k + \tau_{j/k})] = \\ &= E[\eta_k^{\,2}] + E[\eta_k \tau_{j/k}] + E[\eta_k \tau_{i/k}] + E[\tau_{i/k} \tau_{j/k}] \quad \forall i, j \in I_k \end{aligned}$$

Because all the variables η_k, $\tau_{i/k}$ and $\tau_{j/k}$, have zero mean and are mutually independent, the first term is equal to the variance of η_k and the others are zero being the covariances of independent random variables:

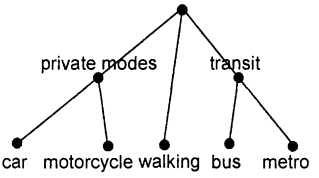
$$Cov[\varepsilon_i, \varepsilon_j] = Var[\eta_k] \qquad \forall i, j \in I_k \tag{3.3.21}$$

However, if two elementary alternatives i and j belong to different groups (h and k) all the terms are zero as is the covariance between ε_i and ε_j.

The variance of η_k can be expressed as a function of the two parameters θ and θ_o:

$$Var[\eta_k] = Var[\varepsilon_j] - Var[\tau_{j/k}] = \frac{\pi^2(\theta_o^2 - \theta^2)}{6} \qquad \forall k \tag{3.3.22}$$

From the previous results, the variance-covariance matrix of random residuals has a block diagonal structure. The elements of the main diagonal are all equal to the variance of residuals ε_j expressed by (3.3.20). The covariance between each pair of alternatives belonging to the same group is constant and equal to the value given by equations (3.3.21) and (3.3.22), while the covariance between alternatives belonging to different groups is null. Therefore, if the alternatives of each group are ordered sequentially, the resulting variance-covariance matrix has a block diagonal structure. Fig. 3.3.5 shows a choice tree and the corresponding variance-covariance matrix.



$$\begin{array}{l} \\ \text{car} \\ \text{motorcycle} \\ \text{walking} \\ \text{bus} \\ \text{metro} \end{array} \frac{\pi^2}{6} \begin{array}{c} \begin{array}{ccccc} \text{car} & \text{motorcycle} & \text{walking} & \text{bus} & \text{metro} \end{array} \\ \begin{bmatrix} \theta_o^2 & \theta_o^2 - \theta^2 & 0 & 0 & 0 \\ \theta_o^2 - \theta^2 & \theta_o^2 & 0 & 0 & 0 \\ 0 & 0 & \theta_o^2 & 0 & 0 \\ 0 & 0 & 0 & \theta_o^2 & \theta_o^2 - \theta^2 \\ 0 & 0 & 0 & \theta_o^2 - \theta^2 & \theta_o^2 \end{bmatrix} \end{array}$$

Fig. 3.3.5 Choice tree and variance-covariance matrix of a Single-Level Hierarchical Logit model.

It is also possible to express the correlation coefficient between two alternatives as a function of the parameters introduced:

$$\rho_{ij} = \begin{cases} \dfrac{Cov[\varepsilon_i \varepsilon_j]}{Var[\varepsilon_i]^{1/2} Var[\varepsilon_j]^{1/2}} = \dfrac{\theta_o^2 - \theta^2}{\theta_o^2} = 1 - \delta^2 & \text{if } i, j \in I_k \\ 0 & \text{otherwise} \end{cases} \tag{3.3.23}$$

The parameters θ, θ_o and δ, play a major role in the structure of the Hierarchical Logit model and influence choice probabilities.

First, the parameter δ defined by equation (3.3.18) can take on values in the interval [0,1]. In fact, it is defined by the ratio between two non-negative quantities and, since the variance of ε_j ($\pi^2\theta_o^2/6$) cannot be inferior to that of one of its components $\tau_{j/k}$ ($\pi^2\theta^2/6$) it must be:

$$\theta_o \geq \theta \rightarrow 0 \leq \delta \leq 1$$

As the variance of $\tau_{j/k}$ tends to that of ε_j, i.e. θ tends to θ_o, the parameter δ will tend to one. In this case, the variance of η_k given by (3.3.22) will tend to zero as the covariance (3.3.21) between two alternatives belonging to the same group and the Hierarchical Logit model (3.3.19) reduces to the Multinomial Logit model.

In fact, for δ=1 in (3.3.19) we get:

$$p[j] = \frac{exp(V_j/\theta)}{\sum_{i\in I_k} exp(V_i/\theta)} \cdot \frac{exp\left[ln\sum_{i\in I_k} exp(V_i/\theta)\right]}{\sum_h exp\left[ln\sum_{i\in I_h} exp(V_i/\theta)\right]} = \frac{exp(V_j/\theta)}{\sum_h \sum_{i\in I_h} exp(V_i/\theta)} \quad (3.3.24)$$

which is a Multinomial Logit model with a different expression of the summation at the denominator.

If the variance of $\tau_{j/k}$ tends to zero, i.e. θ tends to zero, the parameter δ will also tend to zero. In this case the two probabilities in the model (3.3.19) will be modified as follows:

- the conditional choice of an elementary alternative within a group degenerates into a deterministic choice of the maximum utility alternative:

$$\lim_{\theta\to 0} p[j/k] = \lim_{\theta\to 0} \frac{exp(V_j/\theta)}{\sum_{i\in I_k} exp(V_i/\theta)} = \begin{cases} 1 & \text{if } V_j = max_{i\in I_k}(V_i) \\ 0 & \text{otherwise} \end{cases} \quad (3.3.25)$$

- the systematic utilities of alternative groups, equal to θY_k, assume the value of the maximum systematic utility among the elementary alternatives in each group:

$$\lim_{\theta\to 0} \theta Y_k = \lim_{\theta\to 0} \theta ln \sum_{i\in I_k} exp(V_i/\theta) = max_{i\in I_k}(V_i)$$

The choice probability of the group therefore becomes:

$$p[k] = \frac{exp[max_{i \in I_k}\{V_i\}/\theta_o]}{\sum_h exp[max_{i \in I_h}\{V_i\}/\theta_o]} \tag{3.3.26}$$

Thus, if the parameter δ is zero, random residuals associated with the conditional utilities of elementary alternatives within a group are zero ($Var[\tau_{j/k}]=0$), the choice between groups is obtained by comparing the alternatives of maximum systematic utility within each group with a probabilistic Logit model, since a random residual at the group level still exists, while the maximum utility alternative is deterministically chosen within each group.

Some special cases of the model presented can be analyzed. If a group k consists of a single alternative j, then $p[j/k]=1$ and the general expression (3.3.19) for this alternative becomes:

$$p[j] = \frac{exp(V_j/\theta_o)}{exp(V_j/\theta_o) + \sum_{h \neq k} exp(\delta Y_h)} \tag{3.3.27}$$

In some applications of the Single-Level Hierarchical Logit model, and in particular for systems of partial shares models covered in the next chapter, the systematic utility of alternative j, V_j, is decomposed into two parts: a group-specific systematic utility, V_k, and the alternative-specific systematic utility relative to each alternative j, $V_{j/k}$:

$$V_j = V_k + V_{j/k} \tag{3.3.28}$$

This formulation leads to an alternative formulation of choice probabilities $p[j/k]$ and $p[k]$. By replacing (3.3.28) in (3.3.12) and (3.3.17) respectively it follows:

$$p[j/k] = \frac{exp(V_j/\theta)}{\sum_{i \in I_k} exp(V_i/\theta)} = \frac{exp[(V_k + V_{j/k})/\theta]}{exp(V_k/\theta) \cdot \sum_{i \in I_k} exp(V_{i/k}/\theta)} = \frac{exp(V_{j/k}/\theta)}{\sum_{i \in I_k} exp(V_{i/k}/\theta)} \tag{3.3.29}$$

and

$$p[k] = \frac{exp(V_k/\theta_o + \delta Y_k')}{\sum_h exp(V_h/\theta_o + \delta Y_h')} \tag{3.3.30}$$

because:

$$\delta Y_k = \delta ln \sum_{j \in I_k} exp(V_j/\theta) = \delta ln \sum_{j \in I_k} exp[(V_k + V_{j/k})/\theta] =$$

$$= \delta ln \left[exp(V_k/\theta) \cdot \sum_{j \in I_k} exp(V_{j/k}/\theta) \right] = \delta V_k/\theta + \delta ln \sum_{j \in I_k} exp(V_{j/k}/\theta) =$$

$$= V_k/\theta_o + \delta Y_k'$$

where Y'_k is the logsum variable of group k obtained with the alternative specific systematic utilities $V_{j|k}$.

3.3.3. The Multi-Level Hierarchical Logit model*

The Single-Level Hierarchical Logit model described in the previous section is a first generalization of the Multinomial Logit model. However, it retains many simplifying assumptions such as the constancy of covariance between the alternatives belonging to each group and a single level of correlation, or grouping, of alternatives. These assumptions can be generalized considerably as described in the following. The starting point is once again the representation of the choice process and of the covariance between the perceived utilities by means of a general choice tree, from which the name "Tree Logit", sometimes given to these models, derives.

The leaves, or terminal nodes, of the tree correspond to elementary choice alternatives (e.g. different transport modes). Nodes i, j, l in Fig. 3.3.6 are elementary alternatives belonging to the total choice set I. Each intermediate node r can be seen as representing a conditional choice in which the decision-maker chooses from a set of available elementary and/or compound alternatives corresponding to the leaves and/or intermediate nodes directly linked to node r. Thus, each intermediate node represents a compound alternative, i.e. the set of elementary alternatives which can be reached by the intermediate node itself. In conclusion, at each intermediate node the choice is made among all the elementary alternatives, which can be reached, directly or indirectly, through other intermediate nodes, from the node itself. In the example in Fig. 3.3.6, the choice represented by node r is made between alternatives i, j, l, with the elementary alternatives i and l grouped in the compound alternative f.

More formally, the following elements in Fig. 3.3.6 can be defined on the choice tree:

o is the *root or initial node*, the beginning of the decision process;

i, j, l are the *terminal nodes or leaves*, the elementary choice alternatives;

r is the *generic node* of the tree; if this is an *intermediate* (or *structural*) node, it represents both a group of alternatives (compound alternative) and an intermediate choice;

I is the set of elementary alternatives or choice set;

I_r is the set of *descendant nodes (children) of r*; the set of nodes which can be reached directly by r; it represents the set of elementary or compound choice alternatives available for the conditional choice in r; $I_r = \varnothing$, if $r \in I$;

$a(r)$ is the first ancestor of node r, or a node linked to r by the single oriented link $(a(r),r)$ belonging to the graph, $a(o)=\varnothing$;

A_r is the set of ancestors of r; set of nodes belonging to the only route linking the root o and r, excluding r and the root o, $A_r \equiv \{a(r), a(a(r))\ldots\}$;

$p(r,s)$ is the first common ancestor of the pair of nodes r and s.

In this formalism single nodes are indicated with lower/case letters (o, i, j, l, r, s), groups of nodes with capital letters (A, I), and the generic node related to a particular node with lower/case-letter functions of the node itself, $a(r)$, $p(r,s)$.

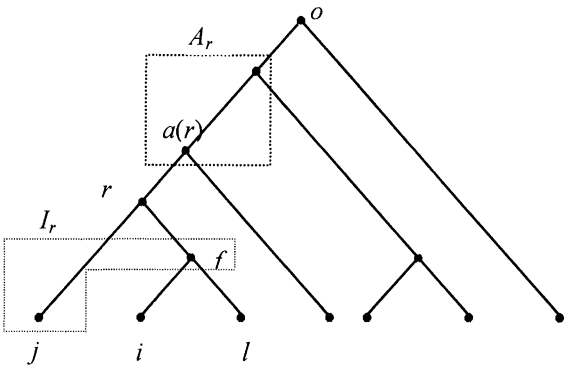


Fig. 3.3.6 Choice tree of Multi-Level Hierarchical Logit models.

At each choice node, whether intermediate or initial, it is assumed that a conditional choice is made among all the available alternatives. These are represented by nodes r, and may be either elementary alternatives (leaves of the tree) or compound alternatives (intermediate nodes). The node representing the choice will be $a(r)$ and the set of choice alternatives will be $I_{a(r)}$

To simulate conditional choice a perceived utility $U_{r/a(r)}$ is assigned to each node (alternative) r. This is a random variable which, as usual, is decomposed into the sum of its mean, V_r, and of a random residual, $\varepsilon_{r/a(r)}$, with the following properties:

- V_r is the expected value of the perceived utility $U_{r/a(r)}$ if r is a leaf of the tree. If r is an intermediate node, V_r is the expected value of the maximum perceived utility (EMPU or inclusive variable) for the alternatives, whether elementary or not, belonging to I_r;
- the random residuals $\varepsilon_{r/a(r)}$ of all nodes r descendants of $a(r)$ are assumed to be i.i.d. Gumbel variables with null mean and parameter $\theta_{a(r)}$. Therefore, the variance $Var[\varepsilon_{r/a(r)}]=\pi^2\theta^2_{a(r)}/6$ is associated with the conditional choice made at node $a(r)$ from all the elementary alternatives, directly or indirectly reached from $a(r)$.

From the above assumptions it results:

$$\begin{aligned} &U_{r/a(r)} = V_r + \varepsilon_{r/a(r)} \quad \forall r \in I_{a(r)} \\ &E[\varepsilon_{r/a(r)}] = 0 \\ &Var[\varepsilon_{r/a(r)}] = \frac{\pi^2\theta^2_{a(r)}}{6} \end{aligned} \tag{3.3.31}$$

From the results on the expected value of the maximum of Gumbel variables referred to in section 3.3.1, the systematic utility assigned to any node can be determined recursively by starting from the leaves as:

$$V_r = \begin{cases} E[U_{r/a(r)}] & \text{if } r \in I \\ \theta_r ln \sum_{h \in I_r} exp(V_h / \theta_r) = \theta_r Y_r & \text{if } r \notin I \end{cases} \quad (3.3.32)$$

Under the above hypotheses, the conditional probability of choosing alternative r at the choice node $a(r)$ is expressed by a Multinomial Logit model:

$$p[r / a(r)] = \frac{exp(V_r / \theta_{a(r)})}{\sum_{r' \in I_{a(r)}} exp(V_{r'} / \theta_{a(r)})} \quad (3.3.33)$$

and also, for (3.3.32):

$$p[r / a(r)] = \frac{exp(V_r / \theta_{a(r)})}{exp(Y_{a(r)})} \quad (3.3.34)$$

If the alternative r is a compound alternative (i.e. r is an intermediate node) for (3.3.32), the numerator of (3.3.33) becomes:

$$exp(V_r / \theta_{a(r)}) = exp\left(\frac{\theta_r}{\theta_{a(r)}} \; Y_r \right) = exp(\delta_r Y_r)$$

where δ_r is the ratio of coefficients θ_r and $\theta_{a(r)}$. It is analogous to the coefficient δ introduced in the previous section (see equation 3.3.18) and, as such, must be included in the interval [0,1]. Expression (3.3.33) and (3.3.34) can be reformulated as:

$$p[r / a(r)] = \frac{exp(\delta_r Y_r)}{\sum_{r'} exp(V_{r'} / \theta_{a(r)})} = \frac{exp(\delta_r Y_r)}{exp(Y_{a(r)})} \quad (3.3.35)$$

Finally, the absolute (unconditional) probability of choosing the elementary alternative $j \in I$, can be obtained from the definition of conditional probability and from the assumptions made on the tree choice mechanism:

$$p[j] = p[j/a(j)] \cdot p[a(j)/a(a(j))] \cdot \ldots \quad j \in I$$

or

$$p[j] = p[j/a(j)] \, \Pi_{r \in A_j} \, p[r/a(r)] \quad j \in I \quad (3.3.36)$$

Replacing expression (3.3.34) and (3.3.35) in equation (3.3.36) we get:

$$p[j] = \frac{exp(V_j/\theta_{a(j)})}{exp(Y_{a(j)})} \cdot \Pi_{r \in A_j} \frac{exp(\delta_r Y_r)}{exp(Y_{a(r)})} \qquad j \in I \tag{3.3.37}$$

and also

$$\begin{aligned} p[j] &= \frac{exp(V_j/\theta_{a(j)})}{exp(Y_o)} \cdot \Pi_{r \in A_j} \frac{exp(\delta_r Y_r)}{exp(Y_r)} = \\ &= \frac{exp(V_j/\theta_{a(j)})}{exp(Y_o)} \cdot \Pi_{r \in A_j} exp[(\delta_r - 1)Y_r] \quad j \in I \end{aligned} \tag{3.3.38}$$

Absolute choice probabilities $p[j]$ can therefore be computed recursively through the following steps:

given: θ_r, $r \notin I$ with $\theta_r = 0$ if $r \in I$
I_r $r \notin I$ with $I_r = \varnothing$ if $r \in I$
V_j $\forall j \in I$

- calculate $\delta_r = \theta_r / \theta_{a(r)}$ for each node r;
- recursively calculate values Y_r, with expression (3.3.32);
- calculate probabilities $p[j], j \in I$, with expression (3.3.38).

The model described can be demonstrated with the choice tree in Fig. 3.3.7. The leaves of the tree (*AI*, *CD*, *CP*, *BS*, *ST*, *FT*) represent the elementary choice alternatives which, in this example, are the transport modes available for an intercity trip: air (*AI*), car driver (*CD*), car passenger (*CP*), bus (*BS*), slow train (*ST*) and fast train (*FT*). The intermediate nodes represent groups of alternatives, or compound alternatives. Node *CR* represents the car, combining the two alternatives of car driver and car passenger, node *LT* public land transport modes (bus, slow train and fast train), while node *RW* combines the railway alternatives. Finally, the respective values of parameters θ and δ are assigned to each intermediate node and to the root.

Following expression (3.3.36), the choice probability of fast train (*FT*) can be written as:

$$p[FT] = p[FT/RW] \cdot p[RW/LT] \cdot p[LT/o]$$

where

$$p[FT/RW] = \frac{exp(V_{FT}/\theta_{RW})}{[exp(V_{ST}/\theta_{RW}) + exp(V_{FT}/\theta_{RW})]} = \frac{exp(V_{FT}/\theta_{RW})}{exp(Y_{RW})}$$

with

$$Y_{RW} = ln[exp(V_{ST}/\theta_{RW}) + exp(V_{FT}/\theta_{RW})]$$

$$\begin{aligned} p[RW/LT] &= \frac{exp(\theta_{RW} Y_{RW}/\theta_{LT})}{exp(\theta_{RW} Y_{RW}/\theta_{LT}) + exp(V_{BS}/\theta_{LT})} = \\ &= \frac{exp(\delta_{RW} Y_{RW})}{exp(\delta_{RW} Y_{RW}) + exp(V_{BS}/\theta_{LT})} = \frac{exp(\delta_{RW} Y_{RW})}{exp(Y_{LT})} \end{aligned}$$

with

$$Y_{LT} = ln[exp(\delta_{RW} Y_{RW}) + exp(V_{BS} / \theta_{LT})] \tag{3.3.39}$$

$$p[LT / o] = \frac{exp(\theta_{LT} Y_{LT} / \theta_o)}{exp(\theta_{LT} Y_{LT} / \theta_o) + exp(\theta_{CR} Y_{CR} / \theta_o) + exp(V_{AI} / \theta_o)} =$$

$$= \frac{exp(\delta_{LT} Y_{LT})}{[exp(\delta_{LT} Y_{LT}) + exp(\delta_{CR} Y_{CR}) + exp(V_{AI} / \theta_o)]} = \frac{exp(\delta_{LT} Y_{LT})}{exp(Y_o)}$$

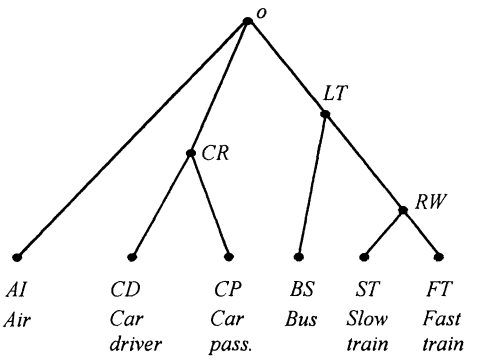
with

$$Y_{CR} = ln[exp(V_{CD} / \theta_{CR}) + exp(V_{CP} / \theta_{CR})]$$

$$Y_o = ln[exp(\delta_{LT} Y_{LT}) + exp(\delta_{CR} Y_{CR}) + exp(V_{AI} / \theta_o)]$$

The absolute choice probability can be written in the form (3.3.38) as follows:

$$P[RW] = \frac{exp(V_{FT} / \theta_{RW})}{exp(Y_o)} \cdot exp[(\delta_{LT} - 1) Y_{LT}] \cdot exp[(\delta_{RW} - 1) Y_{RW}]$$



$$\sum = \frac{\pi^2}{6} \begin{array}{c} \\ AI \\ CD \\ CP \\ BS \\ ST \\ FT \end{array} \begin{array}{c} \begin{array}{cccccc} AI & CD & CP & BS & ST & FT \end{array} \\ \begin{bmatrix} \theta_o^2 & & & & & \\ & \theta_o^2 & \theta_o^2 - \theta_{CR}^2 & & & \\ & \theta_o^2 - \theta_{CR}^2 & \theta_o^2 & & & \\ & & & \theta_o^2 & \theta_o^2 - \theta_{LT}^2 & \theta_o^2 - \theta_{LT}^2 \\ & & & \theta_o^2 - \theta_{LT}^2 & \theta_o^2 & \theta_o^2 - \theta_{RW}^2 \\ & & & \theta_o^2 - \theta_{LT}^2 & \theta_o^2 - \theta_{RW}^2 & \theta_o^2 \end{bmatrix} \end{array}$$

Fig. 3.3.7 Choice tree and variance-covariance matrix for a Multi-Level Hierarchical Logit model.

This choice probability can be seen as resulting from a choice process in which the decision-maker first chooses the compound alternative "collective land transport" represented by node LT from the available alternatives, which in this case

are air, the compound alternative "car" and the compound alternative "collective land transport". Subsequently, he/she chooses the group "train" from the alternatives available within the land transport group (bus and train), and finally fast train from the two elementary alternatives (fast and slow train) which make up the train group.

Returning to the general model, it is possible to express variances and covariances of the random residuals as a function of the parameters θ_r . Rigorous demonstration of these results involves the use of GEV models described in section 3.3.5. The same results can be (approximately) obtained by using the total variance decomposition method described for the Single-Level Hierarchical Logit model in the previous section. It is assumed that the total variance of all the alternatives is constant and equal to:

$$Var[\varepsilon_j] = \pi^2 \theta_o^2/6 \tag{3.3.40}$$

The total random residual of each elementary alternative ε_j is decomposed into the sum of independent zero mean random variables $\tau_{a(r),r}$, associated with each link of the choice tree. Therefore the total variance of an elementary alternative is equal to the sum of the variances corresponding to the links of the (single) route connecting the root to the leaf representing it. Furthermore it is assumed that the variance of random residuals for all the elementary alternatives j reached from any intermediate node r and associated to the conditional choice represented by r itself, is constant and equal to $\pi^2\theta_r^2/6$. It follows that for all these alternatives, the sum of the contributions of the variances relative to the links which connect r to j, must be constant and equal to $\pi^2\theta_r^2/6$:

$$Var[\varepsilon_{j/r}] = \pi^2\theta_r^2/6 = Var[\tau_{a(j),j}] + Var[\tau_{a(a(j)),a(j)}] + \ldots + Var[\tau_{r,f(r,j)}]$$

where $f(r, j)$ is the only descendant of r that is on the route from r to j. Therefore, in the example in Fig. 3.3.7, the variance of the elementary alternatives *BS*, *ST* and *FT* corresponding to the conditional choice between collective land transport modes represented by intermediate node *LT* is constant and equal to $\pi^2\theta_{LT}^2/6$. This variance will correspond to the fraction of variance associated to the link (*LT*, *BS*) and to the sum of the variances associated with links (*LT*, *RW*) and (*RW*, *ST*) or to the links (*LT*, *RW*) and (*RW*, *FT*). The variance of the random residuals of the elementary alternatives relative to the conditional choice represented by node $a(r)$, predecessor of r, is in turn the sum of the variance corresponding to r and the non-negative term, $Var[\tau_{a(r),r}]$, associated with link ($a(r)$, r); this variance will therefore not be inferior to that associated with r, or:

$$\theta_{a(r)} \geq \theta_r \tag{3.3.41}$$

The variance contribution associated with each link ($a(r)$, r) of the graph can be expressed as:

$$Var[\tau_{a(r),r}] = \frac{\pi^2}{6}(\theta_{a(r)}^2 - \theta_r^2) \tag{3.3.42}$$

Inequality (3.3.41) can be generalized, assigning null variance and $\theta_j = 0$ to the leaves of the graph, it yields:

$$\theta_j \leq \theta_{a(j)} \leq \ldots \leq \theta_o \tag{3.3.43}$$

From the preceding expression and the definition of the coefficients $\delta_r = \theta_r/\theta_{a(r)}$, it follows that these coefficients must belong to the interval [0,1].

Continuing with the example in Fig. 3.3.7, the variance of alternatives *ST* and *FT* relative to the conditional choice between railway services (node *RW*) will be $\pi^2\theta_{RW}^2/6$, while that relative to the choice between collective land transport modes (node *LT*) will be $\pi^2\theta_{LT}^2/6$ with $\theta_{LT} \geq \theta_{RW}$; the variance contribution assigned to link (*LT*, *RW*) will be $\pi^2(\theta_{LT}^2 - \theta_{RW}^2)/6$.

The variance decomposition model described allows to derive the covariances between any two elementary alternatives *i* and *j*. This covariance will correspond to the sum of the variances of random residuals $\tau_{a(r),r}$ (which are independent with zero mean) associated with the links common to the two routes connecting the root to leaves *i* and *j*. Because of the tree structure, these routes can have in common only links from the root to the first separation node, which coincides with the last node in common. By applying equation (3.3.42) repeatedly, the covariance of ε_i and ε_j will be:

$$Cov[\varepsilon_i;\varepsilon_j] = \frac{\pi^2(\theta_o^2 - \theta_{p(i,j)}^2)}{6} \qquad \forall i,j \in I \tag{3.3.44}$$

where $p(i,j)$ is the first common ancestor to elementary nodes *i* and *j*.

If two alternatives have the root node as their first common ancestor, i.e. they do not belong to any intermediate compound alternative, their covariance is zero. The correlation coefficient between two elementary alternatives can be deduced from expression (3.3.40) and (3.3.44) as follows:

$$\rho[i,j] = \frac{Cov[\varepsilon_i;\varepsilon_j]}{[Var[\varepsilon_i]\cdot Var[\varepsilon_j]]^{1/2}} = \frac{\theta_o^2 - \theta_{p(i,j)}^2}{\theta_o^2} = 1 - \frac{\theta_{p(i,j)}^2}{\theta_o^2} \tag{3.3.45}$$

For the tree in Fig. 3.3.7, the covariance between alternatives *ST* and *FT* is given by $\pi^2(\theta_o^2 - \theta_{RW}^2)/6$, the sum of the variances relative to links (*o*, *LT*) and (*LT*, *RW*). The covariance between *ST* and *BS* will be $\pi^2(\theta_o^2 - \theta_{LT}^2)/6$ which, as stated before, is less than or equal to the covariance between *FT* and *ST*. In the literature, the parameter θ_o is sometimes taken to be equal to one since, as will be seen in Chapter 8 on transport demand estimation, only parameters δ_r can be estimated. Since all the parameters θ_r but one can be obtained from coefficients δ_r, setting θ_o=1 allows to express all the other parameters. In this case, the covariance and the correlation coefficient between any two elementary alternatives become respectively:

$$Cov[\varepsilon_i, \varepsilon_j] = \frac{\pi^2 (1 - \theta^2_{p(i,j)})}{6}$$

$$\rho[\varepsilon_i, \varepsilon_j] = 1 - \theta^2_{p(i,j)}$$

In conclusion, the structure of the choice tree is also the structure of the covariances between elementary alternatives. Two alternatives that have no nodes in common along the route connecting them to the root o are independent. On the other hand, covariance between elementary alternatives i and j belonging to the same group (their routes meet at an intermediate node) is larger the "further" their first common ancestor is from the root node and the smaller the parameter $\theta_{p(i,j)}$ associated with this node. Furthermore, the covariance between the perceived utility of two alternatives i and j whose first common ancestor ($p(i,j)=a(i)=a(j)$) coincides with their ancestors is not less than the covariance each of them has with any other alternative. Continuing with the example of Fig. 3.3.7, the covariance between *ST* and *FT* will be greater than or equal to that of each of the two elementary alternatives with any other elementary alternative.

Choice probabilities are significantly affected by the values of parameters θ_r, and therefore by the levels of correlation between alternatives. Fig. 3.3.8 shows the values of choice probabilities for the alternatives in Fig. 3.3.7, for different parameters θ_r and assuming that all systematic utilities have the same value: $V_{AI}=V_{CD}=V_{CP}=V_{BS}=V_{ST}=V_{FT}$. If the alternatives are independent (specification nr.1 $\theta_r/\theta_o = 1\ \forall r$), the model becomes a Multinomial Logit and all the alternatives have equal choice probabilities. As the correlation increases, i.e. as parameters θ_{CR}, θ_{LT}, and θ_{RW} decrease, the choice probability of the most correlated alternatives tends to decrease. For example in specification nr. 3, the alternatives belonging to the two groups car (*CD*, *CP*) and collective land transport (*BS*, *ST*, *FT*) are strongly correlated with a correlation coefficient $\rho = 0.9775$. They tend to be seen as a single alternative and their choice probabilities tend to be the equal shares of the probability of a single alternative associated with each group. For the same reasons, the choice probability of alternative *AI*, which is not correlated with any other alternative, is larger the larger the correlation of the alternatives belonging to the various groups (specifications nr. 2 and 3).

From the previous results it can easily be demonstrated that Multinomial Logit and Single-Level Hierarchical Logit models are special cases of the Multi-Level Hierarchical Logit. Two different approaches can be used for the Multinomial Logit model. In the first approach, the tree is that of the Multinomial Logit model described in Fig. 3.3.1. In this case, there are no intermediate nodes and the ancestor $a(j)$ of any leaf $j \in I$ is the root o, it then results $\theta_{a(j)}=\theta_o$, $A_j = \varnothing$ and by applying expression (3.3.38) it follows:

$$p[j] = \frac{exp(V_j / \theta_o)}{exp(Y_o)}$$

which, by developing the term $exp(Y_o)$, gives rise to the expression (3.3.6) of the Multinomial Logit.

SPECIFICATION NR.	1	2	3	4	5	6	7
θ_{LH}/θ_o	1.000	0.900	0.150	1.000	1.000	0.800	0.400
θ_{CR}/θ_o	1.000	0.900	0.150	0.800	0.800	0.600	0.200
θ_{RW}/θ_o	1.000	0.900	0.150	0.600	0.200	0.600	0.200
p[AI]	0.166	0.180	0.304	0.190	0.205	0.212	0.280
p[CD]	0.166	0.168	0.169	0.166	0.178	0.161	0.161
p[CP]	0.166	0.168	0.169	0.166	0.178	0.161	0.161
p[BS]	0.166	0.161	0.120	0.190	0.205	0.174	0.165
p[FT]	0.166	0.161	0.120	0.144	0.117	0.146	0.117
p[ST]	0.166	0.161	0.120	0.144	0.117	0.146	0.117

Fig. 3.3.8 Choice probabilities of the Multi-Level Hierarchical Logit model of Fig. 3.3.7 for varying parameters.

Alternatively the Multinomial Logit model can be obtained from a tree of any form in which the parameters θ_r of all the intermediate nodes are the same and equal to θ_o. In this case from equation (3.3.44) it results that the covariance between any pair of alternatives is equal to zero (independent residuals), the coefficients $\delta_r=\theta_r/\theta_{a(r)}$ are all equal to one, and (3.3.38) reduces to the MNL expression.

The Single-Level Hierarchical Logit model described in the previous section can be considered as a special case of a tree, which has only one level of intermediate nodes

$$a(a(j))=o \quad \forall j \in I$$

Furthermore, parameters θ_r are all equal to θ while the parameter associated with the root is still indicated with θ_o. It can easily be demonstrated that the choice probability (3.3.19) obtained for the Single-Level Hierarchical Logit model results as a special case of expression (3.3.38).

Finally, as in the case of Single-Level Hierarchical Logit model, a systematic utility can be assigned to structural or intermediate nodes. This could be the part of the systematic utility common to all the alternatives connected by an intermediate node. In this case, if r is a structural node and V_r the systematic utility assigned to it, equation (3.3.35) becomes:

$$p[r/a(r)] = \frac{exp(V_r/\theta_{a(r)} + \delta_r Y'_r)}{exp(Y_{a(r)})}$$

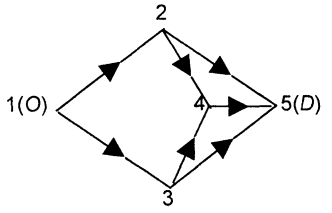
where Y'_r is the logsum variable associated with a node r calculated without the systematic utility V_r, "transferred" to the structural node. Specifications of this type will be used in Chapter 4.

3.3.4. The Cross-Nested Logit model*

The Cross-Nested Logit model can be seen as a generalization of the Hierarchical Logit model allowing a non block-diagonal structure of the variance-covariance matrix. In this model an alternative may belong to more than one group, or nest, with different degrees of membership[13].

As an example, the path choice context reported in Fig. 3.3.9 can be considered. There are four alternatives (paths *A*, *B*, *C*, *D*). It can be assumed that there is a covariance between the perceived utilities of paths *A* and *B* (link (1,2) in common), between paths *B* and *C* (link (4,5) in common) and between paths *C* and *D* (link (1,3) in common). Such a covariance structure cannot be represented by a tree and, in fact, the variance-covariance matrix does not have the typical block-diagonal structure (see Fig. 3.3.9). Using a Cross-Nested structure, on the contrary, three "cross" nests corresponding to the three assumed binary correlations can be specified. Thus alternative *B* belongs to nests 1 and 2 and alternative *C* belongs to nests 2 and 3 (see Fig. 3.3.10). It should be noted that in the case of Cross-Nested models the graph representing the correlation structure should be referred to as choice graph (it is no longer a tree) even though there is no immediate interpretation as a choice process.

In the choice graph intermediate nodes correspond to a group of alternatives (nest).



Path **A** : (1,2),(2,5)
Path **B** : (1,2),(2,4),(4,5)
Path **C** : (1,3),(3,4),(4,5)
Path **D** : (1,3),(3,5)

	A	B	C	D
A	•	•		
B	•	•	•	
C		•	•	•
D			•	•

variance-covariance matrix

Fig. 3.3.9 Example of path choice and its variance-covariance matrix.

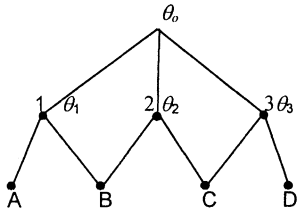


Fig. 3.3.10 Cross-Nested correlation structure for the path choice example in Fig. 3.3.9.

With these assumptions, keeping the same formulation of the Single-Level Hierarchical Logit model, the choice probability of the generic alternative *j*, can be expressed as:

$$p[j] = \sum_k p[j/k] \cdot p[k] \tag{3.3.46}$$

where k represents the generic nest in the single level nesting structure. The difference from the Hierarchical model is that the summation is extended over all nests. This to account for the fact that an alternative can belong, in principle, to any nest k. The degree of membership of an alternative j to a nest k is denoted by α_{jk} and is included in the [0-1] interval. Degrees of membership have to satisfy the following normalizing equation:

$$\sum_k \alpha_{jk} = 1 \quad \forall j \tag{3.3.47}$$

The analytical expressions of $p[j/k]$ and $p[k]$ are as follows:

$$p[j/k] = \frac{\alpha_{jk}^{1/\delta_k} e^{V_j/\theta_k}}{\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k}}; \qquad p[k] = \frac{\left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k}\right)^{\delta_k}}{\sum_{k'} \left(\sum_{i \in I_{k'}} \alpha_{ik'}^{1/\delta_{k'}} e^{V_i/\theta_{k'}}\right)^{\delta_{k'}}} \tag{3.3.48}$$

where I_k is the generic set of alternatives belonging to nest k, θ_k is the parameter associated to an intermediate node, θ_o the parameter associated to the root and δ_k the ratio θ_k/θ_o. Combining equations (3.3.46) and (3.3.48) it results:

$$p[j] = \frac{\sum_k \left[\alpha_{jk}^{1/\delta_k} e^{V_j/\theta_k} \cdot \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k - 1} \right]}{\sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k}} \tag{3.3.49}$$

Analogously to the Hierarchical Logit Model, the parameters δ_k reproduce the correlation among the alternatives, and for δ_k=1 (i.e. θ_k=θ_o) $\forall k$, the Multinomial Logit model (3.3.6) derives from equation (3.3.49):

$$p[j] = \frac{\sum_k \alpha_{jk} e^{V_j/\theta_o}}{\sum_k \sum_{i \in I_k} \alpha_{ik} e^{V_i/\theta_o}} = \frac{e^{V_j/\theta_o} \cdot \sum_k \alpha_{jk}}{\sum_i e^{V_i/\theta_o} \cdot \sum_k \alpha_{ik}} = \frac{e^{V_j/\theta_o}}{\sum_i e^{V_i/\theta_o}}$$

The Cross-Nested Logit model can be derived from the general assumptions of random utility theory as a special case of Generalized Extreme Value (GEV) model as shown in Appendix 3A.

The Cross-Nested Logit model can be seen as a model combining multiple Hierarchical Logit models. In fact, any Cross-Nested specification gives rise to a Hierarchical Logit model for each combination of limit values [0/1] of the membership vector α. For example in the cross-nested structure of Fig. 3.3.10, there are twelve coefficients α_{jk} subject to the following constraints:

$$\begin{aligned} \alpha_{A1} + \alpha_{A2} + \alpha_{A3} &= 1 \\ \alpha_{B1} + \alpha_{B2} + \alpha_{B3} &= 1 \\ \alpha_{C1} + \alpha_{C2} + \alpha_{C3} &= 1 \\ \alpha_{D1} + \alpha_{D2} + \alpha_{D3} &= 1 \end{aligned}$$

With the choice graph depicted it has also been implicitly assumed that:

$$\begin{aligned} \alpha_{A2} = \alpha_{A3} = \alpha_{B3} = \alpha_{C1} = \alpha_{D1} = \alpha_{D2} = 0 \\ \alpha_{A1} = \alpha_{D3} = 1 \end{aligned} \qquad (3.3.50)$$

Consequently there are only four unknown parameters α_{B1}, α_{B2}, α_{C2}, α_{C3} and the effective constraints are:

$$\begin{aligned} \alpha_{B1} + \alpha_{B2} &= 1 \\ \alpha_{C2} + \alpha_{C3} &= 1 \end{aligned}$$

There are four different possible combinations of limit values [0/1] of the vector α consistent with the above constraints that are shown in Fig. 3.3.11:

$$\begin{array}{c} \\ \begin{pmatrix} \alpha_{B1} \\ \alpha_{B2} \\ \alpha_{C2} \\ \alpha_{C3} \end{pmatrix} \end{array} \quad \begin{array}{c} 1 \\ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \end{array} \quad \begin{array}{c} 2 \\ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \end{array} \quad \begin{array}{c} 3 \\ \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \end{array} \quad \begin{array}{c} 4 \\ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \end{array}$$

Fig. 3.3.11 Possible combinations of limit values [0/1] of the membership parameters α for the correlation structure of Fig. 3.3.10.

For any of these combinations there is a corresponding tree structure (see Fig. 3.3.12).

For intermediate α values any intermediate combination of these four Nested correlation structures can be reproduced. Variances and covariances corresponding to the Cross-Nested Logit models have been specified to reproduce the results obtained for Hierarchical Logit models as a special case:

$$Var[\varepsilon_i] = \frac{\pi^2\theta_o^2}{6} \cdot \sum_k (\alpha_{ik})^{1/2} \cdot (\alpha_{ik})^{1/2} = \frac{\pi^2\theta_o^2}{6} \cdot \sum_k \alpha_{ik} = \frac{\pi^2\theta_o^2}{6}$$
$$Cov[\varepsilon_i, \varepsilon_j] = \frac{\pi^2\theta_o^2}{6} \cdot \sum_k (\alpha_{ik})^{1/2} \cdot (\alpha_{jk})^{1/2} \cdot (1 - \delta_k^2) \qquad (3.3.51)$$

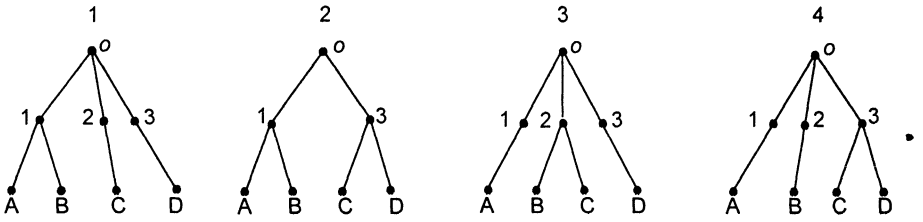


Fig. 3.3.12 Tree correlation structures corresponding to the cross-nested structure of Fig.3.3.10.

Numerical results seem to validate this conjecture even though they haven't been proved formally.

Applying expressions (3.3.51) to the example of Fig. 3.3.9, the following variance-covariance matrix results:

$$\begin{array}{c|cccc|c} & A & B & C & D & \\ A & 1 & (1-\delta_1^2)\alpha_{B1}^{1/2} & 0 & 0 & \\ B & (1-\delta_1^2)\alpha_{B1}^{1/2} & 1 & (1-\delta_2^2)\alpha_{B2}^{1/2} \cdot \alpha_{C2}^{1/2} & 0 & \frac{\pi^2\theta_o^2}{6} \\ C & 0 & (1-\delta_2^2)\alpha_{C2}^{1/2} \cdot \alpha_{B2}^{1/2} & 1 & (1-\delta_3^2)\alpha_{C3}^{1/2} & \\ D & 0 & 0 & (1-\delta_3^2)\alpha_{C3}^{1/2} & 1 & \end{array}$$

The reader can verify that the above matrix gives rise to the variance covariance matrices of the four tree structures of Fig. 3.3.12, when the vector α assumes the corresponding limit values reported in Fig. 3.3.11.

In Fig. 3.3.13 choice probabilities for the example in Fig. 3.3.9 with equal systematic utilities are reported for various hypotheses for the vector α.

α_{B1}	1	0.75	0.5	0.25	0
α_{B2}	0	0.25	0.5	0.75	1
α_{C2}	0	0.25	0.5	0.75	1
α_{C3}	1	0.75	0.5	0.25	0
δ=0.5					
p(A)	0.25	0.2804	0.3039	0.3107	0.2929
p(B)	0.25	0.2196	0.1961	0.1893	0.2071
p(C)	0.25	0.2196	0.1961	0.1893	0.2071
p(D)	0.25	0.2804	0.3039	0.3107	0.2929

Fig. 3.3.13 Choice probabilities for the example in Fig. 3.3.9.

From these results it can be observed that an alternative belonging to several nests has a choice probability lower than another alternative belonging to only one nest with the same systematic utility.

3.3.5. The Generalized Extreme Value (GEV) model*

Generalized Extreme Value models, also known as *GEV* models, are a further generalization of Logit, Hierarchical Logit and Cross-Nested Logit models. Rather than a single model, *GEV* models are a whole class of random utility models. They are defined by a general mathematical formulation including a characteristic function with certain properties; different specifications of the characteristic function give rise to different models such as the models of the Logit family described in previous sections.

GEV models are consistent with the behavioral hypotheses on which random utility theory is based, i.e. that the generic decision-maker associates to each alternative j belonging to his/her choice set a perceived utility. This is decomposed in a deterministic part V_j (systematic utility) and a random residual ε_j. The joint distribution function of random residuals implied by *GEV* models is such that they have the same variance and, in general, non-negative covariances.

A *GEV* model is defined by means of a function $G(y_1, y_2,..., y_m)$ of m variables (m being the number of choice alternatives), continuous and derivable, defined for y_1, $y_2,..., y_n \geq 0$, which has the following properties:

1) $G(.)$ is a non-negative function, $G(.) \geq 0$;
2) $G(.)$ is a homogeneous function of rank $\mu > 0$, that is:
$$G(\alpha y_1, \alpha y_2,..., \alpha y_m) = \alpha^{\mu} G(y_1, y_2,..., y_m);$$
3) $G(.)$ tends asymptotically to infinity for each y_j tending to infinity:
$$lim_{y_j \to \infty} G(y_1, y_2,..., y_m) = \infty \quad j=1, 2,...,m;$$
4) the k^{th} partial derivative of $G(.)$ (or the derivative of rank k of $G(.)$) with respect to a generic combination of k variables y_j, for $j = 1, 2,..., m$, is non-negative if k is odd and non-positive if k is even.

Given a function $G(.)$ satisfying these four properties, the first partial derivative of $G(.)$ with respect to y_j, $\partial G/\partial y_j = G_j(y_1,y_2,\ldots,y_m)$, is homogeneous with rank μ -1, because $G(.)$ is homogeneous of rank μ.

Under the above assumptions it can be shown that choice probabilities of the *GEV* model can be expressed as:

$$p[j] = \frac{y_j \cdot G_j(y_1,...,y_j,...,y_m)}{\mu \cdot G(y_1,...,y_j,...,y_m)} \tag{3.3.52}$$

If y_j is replaced with e^{V_j} (so that the non-negativity of y_j is assured) the *GEV* model can be derived from the hypotheses of random utility theory, assuming that

the joint distribution function $F(\varepsilon)$, of the vector of the random residuals (ε_1, ε_2,..., ε_m) is:

$$F(\varepsilon_1, \varepsilon_2,\ldots, \varepsilon_m) = exp[-G(e^{-\varepsilon_1}, e^{-\varepsilon_2},..., e^{-\varepsilon_m})] \qquad (3.3.53)$$

In fact, as it was seen in section 3.2, the probability of choosing alternative j is equal to:

$$p[j/I] = Pr[V_j - V_k > \varepsilon_k - \varepsilon_j \quad \forall k \neq j, k \in I] \qquad (3.3.54)$$

i.e. the probability that ε_j assumes any value between -∞ and +∞ and that for each alternative $k \neq j$ is $\varepsilon_k < \varepsilon_j + V_j - V_k$. Introducing the joint probability density function of random residuals ε_j, $f(\varepsilon_1, \varepsilon_2,..., \varepsilon_m)$, this probability can also be expressed as:

$$p[j] = \int_{\varepsilon_1=-\infty}^{V_j-V_1+\varepsilon_j} \int_{\varepsilon_2=-\infty}^{V_j-V_2+\varepsilon_j} \ldots \int_{\varepsilon_j=-\infty}^{+\infty} \ldots \int_{\varepsilon_m=-\infty}^{V_j-V_m+\varepsilon_j} f(\varepsilon_1,...,\varepsilon_m)d\varepsilon_1 \ldots d\varepsilon_m \qquad (3.3.55)$$

Alternatively, if $F(\varepsilon_1, \varepsilon_2,..., \varepsilon_m)$, is the cumulated distribution function of random residuals, the partial derivative of F with respect to ε_j, F_j, is equal to the product of the probability density function of ε_j and the joint distribution function for all ε_k with $k \neq j$. The latter, calculated in $\varepsilon_k = V_j - V_k + \varepsilon_j$, gives the probability that each $\varepsilon_k \neq \varepsilon_j$ is less than $V_j - V_k + \varepsilon_j$, for a given value of ε_j. Consequently, equation (3.3.54) can be expressed more synthetically as:

$$p[j] = \int_{\varepsilon_j=-\infty}^{+\infty} F_j(V_j - V_1 + \varepsilon_j,...,\varepsilon_j,...,V_j - V_m + \varepsilon_j)\, d\varepsilon_j \qquad (3.3.56)$$

All the formulations obtained by specifying the joint probability density function $f(\varepsilon_1, \varepsilon_2,..., \varepsilon_m)$, or alternatively the joint probability distribution function $F(\varepsilon_1, \varepsilon_2,..., \varepsilon_m)$, are consistent with the behavioral assumptions of random utility theory expressed by equation (3.3.54).

In particular, the function (3.3.53) where $G(.)$ satisfies the properties 1), 2), 3) and 4) mentioned above, is a cumulated distribution function in that it has the following three properties:

a) $F(.)$ is a non-decreasing function in the ε_j over the whole range of definition;

b) $F(.)$ asymptotically tends to zero if at least one of its variables tends to minus infinity; it tends asymptotically to one if all variables tend to infinity:

$$lim_{\varepsilon_j \to -\infty} F(\varepsilon_1,...,\varepsilon_m) = 0$$
$$lim_{\varepsilon_1,...\varepsilon_m \to +\infty} F(\varepsilon_1,...,\varepsilon_m) = 1$$

c) $F(\cdot)$ is a continuous function from the right.

To demonstrate the first property, it is sufficient to show that the function $G(e^{-\varepsilon_1}, e^{-\varepsilon_2},..., e^{-\varepsilon_m})$, defined earlier is a non-increasing function of ε_j. In fact, from condition 4) on mixed partial derivatives of $G(.)$, it results:

$$G_j(.) \geq 0 \qquad j=1, 2, m \tag{3.3.57}$$

i.e. G(.) is non-decreasing with respect to the variables $e^{-\varepsilon_j}$. It then follows that:

$$\partial G(.)/\partial \varepsilon_j = \partial G(.)/\partial e^{-\varepsilon_j} \cdot \partial e^{-\varepsilon_j}/\partial \varepsilon_j = G_j(.)\cdot(-e^{-\varepsilon_j}) \leq 0$$

The function $G(e^{-\varepsilon_1}, e^{-\varepsilon_2},..., e^{-\varepsilon_m})$ is therefore non-decreasing in $e^{-\varepsilon_j}$ but non-increasing in ε_j.

As for the second property, from equation (3.3.53) and condition 3) required for $G(.)$, it follows:

$$\lim_{\varepsilon_i \to -\infty} F(\varepsilon_1,...,\varepsilon_j,...,\varepsilon_m) = \lim_{\varepsilon_j \to -\infty} exp[-G(e^{-\varepsilon_1},...,e^{-\varepsilon_j},...,e^{-\varepsilon_m})] =$$

$$= exp[-G(e^{-\varepsilon_1},...,\infty,...,e^{-\varepsilon_m})] = exp[-\infty] = 0$$

which is the first of the two limits. The second limit, derives from the homogeneity, condition 2) of $G(.)$ (condition 2) implying that $G(0, 0,\ldots,0)=0$. Therefore from equation (3.3.53) it results:

$$\lim_{\varepsilon_1,...\varepsilon_m \to +\infty} F(\varepsilon_1,...,\varepsilon_m) = \lim_{\varepsilon_1,...\varepsilon_m \to +\infty} exp[-G(e^{-\varepsilon_1},...,e^{-\varepsilon_m})] =$$

$$= exp[-G(0,...,0)] = exp[-0] = 1$$

The third property is easily verified, being $F(.)$ defined by (3.3.53) a continuous function.

Furthermore, it can be demonstrated that the solution of equation (3.3.56) with F defined as in (3.3.53) actually gives the expression (3.3.52) of the choice probabilities defining a GEV model.

In fact, substituting equation (3.3.53) in expression (3.3.56), for the homogeneity of $G(.)$ and $G_j(.)$, it follows:

$$p[j] = \int_{\varepsilon_j=-\infty}^{+\infty} exp[-G(e^{V_1-V_j-\varepsilon_j},...,e^{V_m-V_j-\varepsilon_j})]\cdot G_j(e^{V_1-V_j-\varepsilon_j},...,e^{V_m-V_j-\varepsilon_j})\cdot e^{-\varepsilon_j}\, d\varepsilon_j =$$

$$= \int_{\varepsilon_j=-\infty}^{+\infty} exp\{-[e^{-(V_j+\varepsilon_j)}]^{\mu}\cdot G(e^{V_1},...,e^{V_m})\}\cdot[e^{-(V_j+\varepsilon_j)}]^{\mu-1}\cdot G_j(e^{V_1},...,e^{V_m})\cdot e^{-\varepsilon_j}\, d\varepsilon_j =$$

$$= \int_{\varepsilon_j=-\infty}^{+\infty} \{exp-[e^{-(V_j+\varepsilon_j)}]^{\mu}\}^{G(e^{V_1},...,e^{V_m})}\cdot[e^{-(V_j+\varepsilon_j)}]^{\mu-1}\cdot G_j(e^{V_1},...,e^{V_m})\cdot e^{-\varepsilon_j}\, d\varepsilon_j =$$

$$= \frac{e^{V_j}\cdot G_i(e^{V_1},...,e^{V_m})}{\mu\cdot G(e^{V_1},...,e^{V_m})}\cdot\left|\{exp-[e^{-(V_j+\varepsilon_j)}]^{\mu}\}^{G(e^{V_1},...,e^{V_m})}\right|_{-\infty}^{+\infty} = \frac{e^{V_j}\cdot G_j(e^{V_1},...,e^{V_m})}{\mu\cdot G(e^{V_1},...,e^{V_m})}$$

which is the (3.3.52) with y_j replaced with e^{Vj}.

Multinomial Logit, Single-Level Hierarchical Logit, Multi-Level Hierarchical Logit, and Cross-Nested Logit models can be obtained as special cases of the *GEV* model opportunely specifying the function $G(.)$ as it will be shown in Appendix 3.A.

3.3.6. The Probit model

The Probit model overcomes most of the drawbacks of the Logit model and its generalizations, though at the cost of analytical tractability. It is based on the

hypothesis that residuals ε_j are distributed according to a Multivariate Normal (MVN)[14] random variable with zero mean and general variances and covariances:

$$\begin{aligned} &E[\varepsilon_j] = 0 \\ &Var[\varepsilon_j] = \sigma_j^2 \\ &Cov[\varepsilon_j, \varepsilon_h] = \sigma_{jh} \end{aligned} \tag{3.3.58}$$

Variances and covariances are the elements of the dispersion matrix, Σ, of the random vector ε with a number of row and columns equal to the number of alternatives m. The Multivariate Normal density probability of the vector ε is given by:

$$f(\varepsilon) = [(2\pi)^m det(\Sigma)]^{-1/2} exp[-1/2\varepsilon^T \Sigma^{-1} \varepsilon] \tag{3.3.59}$$

Perceived utilities U_j are also jointly distributed according to a Multivariate Normal with mean vector V and variances and covariances equal to those of residuals ε_j; $U \sim \text{MVN}(V, \Sigma)$.

The choice probability of alternative j can be formally expressed as the joint probability that utility U_j will assume a value within an infinitesimal interval and that the utilities of the other alternatives will have lower values. Clearly this probability must be integrated over all possible values of Uj. This can be expressed formally as (see equation 3.3.55):

$$p[j] = \int_{U_1 < U_j} \cdots \int_{U_j = -\infty}^{\infty} \cdots \int_{U_m < U_j} \frac{exp[-1/2(U-V)^T \Sigma^{-1}(U-V)]}{[(2\pi)^m det(\Sigma)]^{1/2}} dU_1 \ldots dU_m \tag{3.3.60}$$

The Probit model is an additive model if the matrix Σ does not depend on the vector of the systematic utilities V. In this case the choice probability of a generic alternative depends on the differences of systematic utilities. Thus Alternative Specific Attributes (ASA) and their coefficients (ASC) can be replaced by their differences with respect to the value of reference alternative.

To illustrate the effect of variances and covariances on choice probabilities, the case of three alternatives (m = 3), with systematic utilities equal to zero ($V_A = V_B = V_C = 0$) and the following variance-covariance matrix:

$$\sum = \begin{bmatrix} 1 & \sigma_{AB} & 0 \\ \sigma_{AB} & 1 & 0 \\ 0 & 0 & \sigma_C^2 \end{bmatrix}$$

can be considered. Fig. 3.3.14 maps the probability $p[C]$ obtained with the Probit model (3.3.60) for varying values of parameters σ_{AB} and σ_C. As the variance of U_C

increases compared with those of the other alternatives, the choice probability of C increases. In fact, the random residual ε_C becomes dominant over the value of V_C and the perceived utility U_C is either much higher or much lower than the perceived utilities U_A and U_B ($lim_{\sigma_C \to \infty}\ p[C]=0.5$). Also, as the covariance (in this case coincident with the correlation coefficient) between the residuals of alternatives A and B increases, the choice probability of alternative C increases since A and B are increasingly perceived as a single alternative. The same effect was shown in sections 3.3.2 and 3.3.3 for the Hierarchical Logit model.

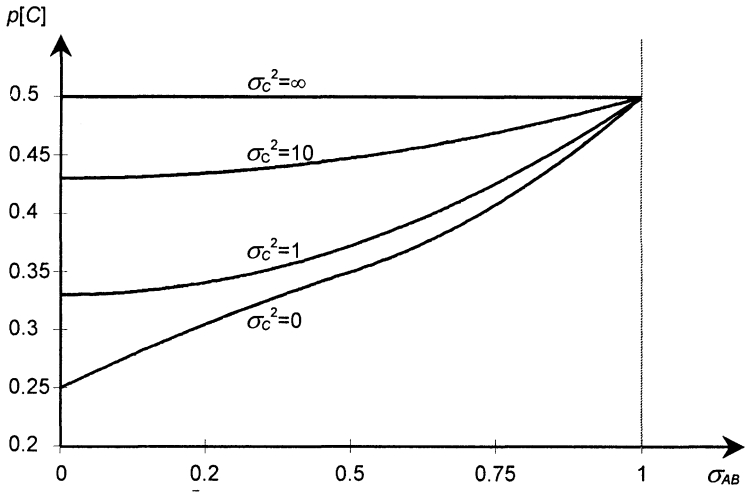


Fig. 3.3.14 Influence of the variance and covariance of residuals on Probit choice probabilities.

In general, the Probit model yields choice probabilities similar to those obtained from Logit and Hierarchical Logit models if the same variance-covariance matrix is assumed. Nevertheless Probit allows a greater flexibility in the specification of this matrix.

Generality of the variance-covariance matrix can also be a problem in the practical use of the Probit model. As a matter of fact in a variance-covariance matrix there are at most $(m(m+1))/2$ different values, where m is the number of choice alternatives. When m is large the specification and calibration of all the possible values can be problematic. Different methods have been proposed to reduce the number of unknown elements in the variance covariance matrix. All these methods assume some structure underlying to the random residuals; the parameters of this structure allow to specify the variance-covariance matrix and are less than all the possible unknowns.

A first method known as *Factor Analytic Probit* expresses the vector of random residuals as a linear function of a vector ζ of independent standard normal variables:

$$\varepsilon_j = \sum_{k=1}^{n} f_{jk}\zeta_k \tag{3.3.61}$$

$$\boldsymbol{\varepsilon} = \boldsymbol{F}\boldsymbol{\zeta} \tag{3.3.62}$$

where:

- $\boldsymbol{\varepsilon}$ is the $(m\times 1)$ vector of multivariate normal distributed random variables (factors) with elements ε_j: $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$;
- $\boldsymbol{F}$ is the $(m\times n)$ matrix of loading with elements f_{ik}, mapping the vector $\boldsymbol{\zeta}$ of standard random variables to the vector $\boldsymbol{\varepsilon}$ of random residuals;
- $\boldsymbol{\zeta}$ is the $(n\times 1)$ vector of identical and independent standard normal distributed random variables with elements ζ_k: $\boldsymbol{\zeta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I})$.

typically $n << m$ and the number of unknown elements is reduced from the $m(m+1)/2$ of the matrix $\boldsymbol{\Sigma}$ to the $m \cdot n$ of the matrix $\boldsymbol{F}$. In the extreme case $(m=n)$ the matrix $\boldsymbol{F}$ is such that:

$$\boldsymbol{F}\boldsymbol{F}^T = \boldsymbol{\Sigma} \tag{3.3.63}$$

and can be obtained through the Cholesky factorization of $\boldsymbol{\Sigma}$ matrix as described subsequently.

From the (3.3.61), the elements of the variance-covariance matrix $\boldsymbol{\Sigma}$ of random residuals ε_j can be expressed as a function of the elements f_{jk} of matrix $\boldsymbol{F}$:

$$Var[\varepsilon_j] = E[\varepsilon_j^2] = E\left[\sum_{k=1}^{n} f_{jk}^2\zeta_k^2\right] = \sum_{k=1}^{n} f_{jk}^2 \cdot E[\zeta_k^2] = \sum_{k=1}^{n} f_{jk}^2 \tag{3.3.64}$$

$$Cov[\varepsilon_j, \varepsilon_h] = E[\varepsilon_j\varepsilon_h] = E\left[\sum_{k=1}^{n} f_{jk}\zeta_k \cdot \sum_{k=1}^{n} f_{hk}\zeta_k\right] = \sum_{k=1}^{n} f_{jk}f_{hk} \cdot E[\zeta_k^2] = \sum_{k=1}^{n} f_{jk}f_{hk} \tag{3.3.65}$$

A relevant application of the factor analytic representation of the Probit model is in path choice as it will be shown in section 4.3.4.1.

This method also simplifies the computation of choice probabilities with simulation methods (Monte Carlo) since most statistic routines draw values from a standard normal distribution rather than vectors of values from a general MVN distribution.

A different method to reduce the number of unknown coefficients in the dispersion matrix is the *Random Coefficients* (*Random Tastes*) *Probit* model. This model is based on the assumptions that the random residual ε_j derives from the dispersion of coefficients β_k over the population of decision makers. In particular each coefficient $\beta^i{}_k$ is assumed equal to an average value β_k plus a random residual $\eta^i{}_k$:

$$\beta^i{}_k = \beta_k + \eta^i{}_k \qquad k=1,2,...,K$$

where K is the total number of coefficients used to define the systematic utilities of the m alternatives. By assuming the η^i_k independently distributed as normal variables with zero mean and variance σ^2_k:

$$\eta^i_k \sim N(0, \sigma^2_k) \quad \forall\, i, k$$
$$Cov[\eta^i_k, \eta^i_h] = 0 \quad \forall\, i, k, h$$

it results:

$$U^i_j = V^i_j + \varepsilon^i_j = \sum_k \beta^i_k X^i_{kj} = \sum_k \beta_k X^i_{kj} + \eta^i_k X^i_{kj}$$

with:

$$V^i_j = \sum_k \beta_k X^i_{kj}; \quad \varepsilon^i_j = \sum_k \eta^i_k X^i_{kj}; \quad \boldsymbol{\varepsilon}^i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$$

$$Var[\varepsilon^i_j] = E\left[\left(\sum_k \eta^i_k X^i_{kj}\right)^2\right] = \sum_k (X^i_{kj}\sigma_k)^2 \tag{3.3.66}$$

$$Cov[\varepsilon^i_j, \varepsilon^i_h] = E\left[\left(\sum_k \eta^i_k X^i_{kj}\right)\cdot\left(\sum_k \eta^i_k X^i_{kh}\right)\right] = \sum_k X^i_{kj} X^i_{kh} \sigma^2_k \tag{3.3.67}$$

where X_{kj} is the value of attribute k in alternative j; it is equal to zero if attribute X_k does not appear in the systematic utility of alternative j. Using this approach the unknown elements of the variance-covariance matrix are reduced (generally) from $(m(m+1))/2$ to K.

The Random Coefficient Probit model can be seen as an application of the factor analytic representation described. In fact, neglecting user index i, it results:

$$\boldsymbol{\varepsilon} = \boldsymbol{X}^T \boldsymbol{\eta} = \boldsymbol{X}^T \boldsymbol{\Sigma}_\eta^{1/2} \boldsymbol{\zeta} = \boldsymbol{F}\boldsymbol{\zeta}$$

where:

- $\boldsymbol{\varepsilon}$ is the $(m\times 1)$ vector of multivariate normal distributed alternative random residual, $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$;
- $\boldsymbol{X}$ is the $(K\times m)$ matrix of attribute values;
- $\boldsymbol{\eta}$ is the $(K\times 1)$ vector of independent normal distributed coefficient random residual, $\boldsymbol{\eta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$;
- $\boldsymbol{\zeta}$ is the $(K\times 1)$ vector of identical and independent standard normal distributed random variables, $\boldsymbol{\zeta} \sim MVN(\mathbf{0}, \mathbf{I})$;
- $\boldsymbol{F}$ is the $(m\times K)$ matrix of loading, $\boldsymbol{F} = \boldsymbol{X}^T \boldsymbol{\Sigma}_\eta^{1/2}$.

and it is immediate to verify that matrix $\boldsymbol{F}$ specified above, introduced in the (3.3.64) and (3.3.65) gives the (3.3.66) and (3.3.67) respectively.

Flexibility of the Probit model is achieved at the cost of computational complexity. In fact, the Probit model does not allow to express analytically choice probabilities since there is no known closed-form solution of the integral (3.3.60).

Numerical integration methods are computationally burdensome when there are more than five alternatives. Calculation of Probit choice probabilities with several alternatives is typically carried out by approximation methods. In the following, two "traditional" methods will be described. However, it should be said that new and more efficient methods are currently being studied.

The *Monte-Carlo simulation* method generates a sample of perceived utilities (these can be thought of as the perceived utilities of a sample of decision-makers) and estimates the choice probability of each alternative j as the fraction of times j is the alternative of maximum perceived utility.

At the k^{th} iteration the method generates:

- a vector $\boldsymbol{\varepsilon}^k = (\varepsilon_1{}^k, \ldots, \varepsilon_m{}^k)^T$ of random residuals extracted from a zero mean Multivariate Normal variable with dispersion matrix $\boldsymbol{\Sigma}$;
- a vector $\boldsymbol{U}^k$ of perceived utilities: $\boldsymbol{U}^k = \boldsymbol{V} + \boldsymbol{\varepsilon}^k$;
- a vector $\boldsymbol{p}^k$ of deterministic alternative choice probabilities: $\boldsymbol{p}^k = (0, \ldots, 1, \ldots, 0)$.

where the value one is associated to the largest component $\boldsymbol{U}^k$ (maximum perceived utility alternative). Consequently, the sample estimate $\hat{p}[j]$ of generic probability $p[j]$ will be:

$$\hat{p}[j] = \frac{1}{n}\sum_{k=1}^{n} p[j/\boldsymbol{\varepsilon}^k] = \frac{n_j}{n} \tag{3.3.68}$$

where $\boldsymbol{\varepsilon}^k$ denotes draw k of vector $\boldsymbol{\varepsilon}$ from a $N(\mathbf{0}, \boldsymbol{\Sigma})$ and n_j is the number of times that alternative j is the maximum perceived utility alternative.

It can be shown that the estimator (3.3.68) is unbiased and efficient. In applications, the vector $\boldsymbol{\varepsilon}$ extracted from a random variable MVN(0,$\boldsymbol{\Sigma}$) can be obtained indirectly by means of m values extracted independently from a standard normal variable $N[0, 1]$. In fact, the positive definite matrix $\boldsymbol{\Sigma}$ can be expressed as the product of a matrix and its transpose:

$$\boldsymbol{\Sigma} = \boldsymbol{C}\,\boldsymbol{C}^T \tag{3.3.69}$$

Matrix $\boldsymbol{C}$ can be obtained, for example, with the Cholesky factorization of matrix $\boldsymbol{\Sigma}$. If z indicates a normal standard vector of dimension $(m \cdot 1)$ the vector $\boldsymbol{C}z$ is distributed according to a Multivariate Normal MVN(0, $\boldsymbol{\Sigma}$):

$$E[\boldsymbol{C}z] = \boldsymbol{C}E[z] = 0$$
$$Var[\boldsymbol{C}z] = E[\boldsymbol{C}zz^T\boldsymbol{C}^T] = \boldsymbol{C}\boldsymbol{I}\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{\Sigma}$$

Thus the vector $\boldsymbol{\varepsilon}$ can be obtained as:

$$\boldsymbol{\varepsilon} = \boldsymbol{C}\,z$$

With the Monte-Carlo method, each extraction can be considered as the execution of a Bernoulli experiment with m possible results. The probability of obtaining the j^{th} result is in fact the choice probability of that alternative $p[j]$. It is therefore possible to estimate the sample variance of the estimate $\hat{p}[j]$ as:

$$Var[\hat{p}[j]] = \frac{1}{n}\hat{p}[j](1 - \hat{p}[j]) \tag{3.3.70}$$

A confidence interval for $p[j]$ can be obtained by assuming that $p[j]$ is approximately distributed as a normal r.v. with mean, $\hat{p}[j]$, given by (3.3.68) and variance given by (3.3.70) for values of n large enough.

In Chapter 7 the Monte-Carlo method will be applied in a different context to calculate Probit path choice probabilities on a network, exploiting the special structure of this problem.

The *Clark approximation* method is based on the moments of a r.v. corresponding to the maximum of normal random variables. The procedure will be illustrated first referring to the choice among three alternatives. In this case perceived utilities are distributed according to a multivariate normal with mean vector $\boldsymbol{V} = (V_1, V_2, V_3)^T$ and the following variance-covariance matrix:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Suppose the choice probability of alternative 3, $p[3]$, has to be computed. Clark's results express the mean V_{12} and the variance S^2_{12} of the random variable $U_{12} = max(U_1, U_2)$ as:

$$\begin{aligned} V_{12} &= V_2 + (V_1 - V_2)\,F(\alpha) + \gamma f(\alpha) \\ S^2_{12} &= var[U_{12}] = m_{12} - V^2_{12} \end{aligned} \tag{3.3.71}$$

where m_{12} is the second moment around zero of the variable U_{12} given by:

$$m_{12} = V^2_2 + \sigma^2_2 + (V^2_1 + \sigma^2_1 - V^2_2 - \sigma^2_2)\,F(\alpha) + (V_1 + V_2)\,\gamma f(\alpha) \tag{3.3.72}$$

The constants γ and α in expression (3.3.71) and (3.3.72) are the standard deviation of the random variable $(U_1 - U_2)$:

$$\gamma = [\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}]^{1/2}$$

and the mean standardized value of the random variable $(U_1 - U_2)$:

$$\alpha = (V_1 - V_2)/\gamma$$

The symbols $f(\alpha)$ and $F(\alpha)$ denote respectively the value of the probability density function and probability distribution function of a normal standard r.v. $N(0,1)$ calculated in α:

$$f(\alpha) = (2\pi)^{-1/2} \exp(-\alpha^2/2)$$

$$F(\alpha) = \int_{-\infty}^{\alpha} f(x)dx$$

Clark's formulas also give the covariance between variables U_j and $U_{12...j-1}$ as:

$$S_{j.12...i} = cov(U_j, U_{12...i}) = \sigma_{ij} + (S_{j.12...i-1} - \sigma_{ij})\, F(\alpha)$$

where it is $i = j\text{-}1$. Thus the covariance between variables U_3 and U_{12} results:

$$S_{3.12} = cov(U_3, U_{12}) = \sigma_{23} + (\sigma_{13} - \sigma_{23})\, F(\alpha)$$

Moreover, if U_{12} is approximated as a normal r.v., it is possible to calculate the probability of choosing alternative 3 as:

$$p[3] = Pr[U_3 \geq U_{12}] = Pr[U_{12} - U_3 \leq 0] \tag{3.3.73}$$

Since the difference between two normal r.v. is still a normal r.v. with a mean given by the difference between the two means V_3 and V_{12} and variance given by the sum of the two variances minus twice the covariance, the choice probability (3.3.73) can be computed as:

$$p[3] = F\left[\frac{V_3 - V_{12}}{(\sigma_3^2 + S_{12}^2 - 2S_{3.12})^{1/2}}\right] \tag{3.3.74}$$

Choice probabilities with more than three alternatives can be calculated by applying sequentially the procedure described. The probability of choosing the generic alternative j can be obtained by computing sequentially the mean variance and covariance of nested pairs of perceived utilities ordered in such a way that j is the last alternative. For example the mean and variance of $U_{12} = max(U_1, U_2)$ as well as its covariance with U_3 are computed first. Subsequently are computed the mean and variance of the variable $U_{123} = max(U_3, U_{12})$, together with its covariance with U_4, and so on until the comparison is made between:

$$U_{12...j-1} = max(U_{j-1}, max(U_{j-2}... max(U_1, U_2)))$$

and U_j to obtain probability $p[j]$ by applying expression (3.3.74). The entire sequence has to be repeated to calculate the probability expression of each alternative.

3.3.7. The Hybrid Logit-Probit model*

The Hybrid Logit-Probit model generalizes Multinomial Logit and Probit models since both of them can be derived as special cases. The Hybrid Logit-Probit model is also computationally more efficient than the Probit model as it will be seen later. The random residual ε_j of the perceived utility of an alternative j is divided into the sum of two independent random variables, ξ_j and υ_j:

$$U_j = V_j + \varepsilon_j = V_j + \xi_j + \upsilon_j \tag{3.3.75}$$

Moreover, the ξ_j are assumed MVN distributed with mean vector $\mathbf{0}$ and generic variance-covariance matrix Σ (capturing the interdependencies among alternatives), while υ_j are i.i.d. Gumbel variables with mean 0 and parameter θ

$$\begin{aligned} \xi &\sim MVN(\mathbf{0}, \Sigma) \\ \upsilon &\sim G(\mathbf{0}, \sigma^2 \boldsymbol{I}) \end{aligned} \tag{3.3.76}$$

In this way, the normal random residuals ξ_j give the general Probit structure to the model while the Gumbel residuals υ_j result in the Logit kernel. As a matter of fact, for a given vector $\bar{\xi}$ the classic Multinomial Logit expression for the choice probability of the generic alternative j results:

$$p[j/\bar{\xi}] = \frac{exp[(V_j + \bar{\xi}_j)/\theta]}{\sum_h exp[(V_h + \bar{\xi}_h)/\theta]} \tag{3.3.77}$$

Consequently the general expression of the choice probability for Hybrid Logit-Probit model is:

$$p[j] = \int_{\xi} p[j/\xi] f(\xi) d\xi \tag{3.3.78}$$

Obviously, the same methods proposed to reduce the number of unknown elements in Σ and to compute choice probabilities (factor analysis, random coefficients, simulation methods, etc.) can still be used for the Hybrid Logit Probit model. For example the variation in utility coefficients and/or in perception of utility attributes can be assumed to give rise to MVN residuals ξ while other factors, such as missing attributes etc, can be assumed to produce residuals υ.

The main advantages of Hybrid Logit Probit with respect to Probit is the possibility to compute numerically choice probabilities in more efficient ways, while keeping a variance-covariance matrix almost as general as the one associated with Probit models[15]. The dimensionality of the multifold integrals involved in both models requires, in most cases, the use of unbiased efficient estimators of choice

probabilities as the MonteCarlo method described in section 3.3.6. The advantage of this model is that it leads to smooth and unbiased choice probability simulator:

$$p[j] = \frac{1}{n}\sum_{k=1}^{n} p[j/\xi^k] \tag{3.3.79}$$

where ξ^k denotes draw k of vector ξ from a MVN($\mathbf{0}$,Σ).

Comparing expression (3.3.79) with its equivalent for the Probit model (3.3.68) it can be observed that, while in the simulation of Probit choice probabilities each draw of vector ξ^k from a multivariate normal yields a vector of deterministic choice probabilities [0,...,1,...,0], in the simulation of Hybrid Logit-Probit choice probabilities, each draw of ξ^k produces a vector of stochastic choice probabilities $[p_1,\ldots, p_j,\ldots, p_n]$ where the generic $p[j]$ is given by equation (3.3.77) with vector $\bar{\xi}$ replaced by vector ξ^k.

3.4. Choice set modeling*

Random utility models simulate the choice made by the generic individual i from a set of alternatives, which make up his/her choice set I^i, under the hypothesis that the analyst is able to specify correctly this set. When this hypothesis is not acceptable, it is necessary to simulate explicitly the composition of the generic decision-maker's choice set. This problem has been tackled by following two basically different approaches. The *implicit approach* simulates the perception/availability of an alternative within the choice model of the alternative. The *explicit approach* simulates explicitly the choice set generation with a specific model.

The first approach has been adopted in many specifications of random utility models proposed in the literature. Some attributes in the systematic utility function of an alternative play the role of "proxy" variables, simulating the availability/perception of that alternative. For example, the number of cars divided by the number of licensed drivers in a household is used to simulate the availability of the car in mode choice models. Attributes with this interpretation can be easily identified in several random utility models described in the next chapter. The implicit approach is undoubtedly simpler from the application point of view, though there is a noticeable lack of consistency since "utility" attributes are mixed with "availability" attributes.

In the explicit approach, the choice probability of an alternative j for decision-maker i is usually expressed through a two-stage choice model:

$$p^i[j] = \sum_{I^i \in G^i} p^i[j, I^i] = \sum_{I^i \in G^i} p^i[j/I^i]p^i[I^i] \tag{3.4.1}$$

where:

I^i is the generic choice set of decision-maker i;

G^i is the set made up of all the possible non-empty choice sets for decision-

maker i (non-empty subsets of the set of all the possible alternatives);

$p^i[j,I^i]$ is the joint probability that decision-maker i will choose the alternative j and that I^i is his/her choice set;

$p^i[j/I^i]$ is the probability that decision-maker i will choose alternative j, his/her choice set being I^i;

$p^i[I^i]$ is the probability that I^i is the choice set of individual i.

The choice probability conditional on set I^i, $p^i[j/I^i]$, can be simulated with one of the random utility models described in section 3.3.

An example of the explicit model of choice set generation can be obtained starting from the general model (3.4.1) and assuming that the probabilities that each single alternative belongs to the choice set are independent of each other:

$$Pr[j \in I^i / h \in I^i] = Pr[j \in I^i] \quad \forall j,h \tag{3.4.2}$$

In this case, the probability $p[I^i]$ can be expressed as :

$$p[I^i] = \frac{\prod_{h \in I^i} p[h \in I^i] \cdot \prod_{k \notin I^i} [1 - p[k \in I^i]]}{1 - p[I^i \equiv \varnothing]} \tag{3.4.3}$$

where the first product is extended to all the alternatives included in I^i and the second to all those not included in I^i; the probability that the choice set is empty is given by:

$$p[I^i \equiv \varnothing] = \prod_j [1 - p[j \in I^i]] \tag{3.4.4}$$

The denominator of expression (3.4.3) normalizes probabilities $p[I^i]$ to take into account the fact that an empty choice set ($I^i \equiv \varnothing$) is usually excluded under the assumption that the decision-maker's choice set includes at least one alternative. Replacing expression (3.4.3) and (3.4.4) in (3.4.1), the choice probability of the generic alternative is:

$$p^i[j] = \frac{\sum_{I^i \in G^i} \left\{ \prod_{h \in I^i} p^i[h \in I^i] \cdot \prod_{k \notin I^i} [1 - p^i[k \in I^i]] \cdot p^i[j / I^i] \right\}}{1 - \prod_j [1 - p^i[j \in I^i]]} \tag{3.4.5}$$

Specification of model (3.4.5) requires a model simulating the probability that generic alternative j belongs to the choice set $p[j \in I^i]$. Various authors have proposed a Binomial Logit model [16]:

$$p[j \in I^i] = \frac{1}{1 + exp\left(\sum_k \gamma_k Y_{kj}^i\right)} \tag{3.4.6}$$

where Y_k are the "availability/perception" variables mentioned above and γ_k are the relative coefficients.

The explicit approach, though very interesting and consistent from a theoretical point of view, poses some computational problems. The number of all the possible choice sets, i.e. the cardinality of G^i, grows exponentially with the number of possible alternatives. This complicates the calculation of choice probabilities (3.4.1), and therefore the joint calibration of the parameters β_k in the systematic utility and γ_k in the choice set model.

An intermediate approach named of Implicit Availability Perception (IAP), is based on the simulation of the availability/perception of each alternative with a model included in the utility function of the random utility model. This approach is based on the generalization of the conventional concepts of availability and choice set. It is assumed that an alternative may have various levels of availability/ perception for the generic decision-maker. It follows that the choice set of a generic decision-maker is seen as a "fuzzy set", allowing for intermediate levels of membership of the single elements in the set. The choice set is no longer represented as a set of Boolean variables [0/1] (1 if the alternative is available/ perceived, 0 otherwise) but as a set of continuous variables $\mu_I(j)$ defined in the interval [0/1]. This can be the case of an alternative theoretically available, but not completely perceived as such for a particular journey, due either to subjective (lack of information, time constraints, state of health, etc.) or to objective (weather conditions, etc.) factors. Obviously, extreme values are still possible, representing respectively non-availability and complete availability and perception of the alternative. The model reproduces different levels of availability/perception of an alternative by directly introducing an appropriate functional transformation of $\mu_I(j)$ into the utility function of the alternative:

$$U_j^i = V_j^i + ln\mu_I^i(j) + \varepsilon_j^i \tag{3.4.7}$$

where

U_j^i is the perceived utility of alternative j for decision-maker i;
V_j^i is the systematic utility of alternative j for decision-maker i;
ε_j^i is the random residual of alternative j for decision-maker i;
$\mu_I^i(j)$ is the level of membership of alternative j to the choice set I^i of decision-maker i ($0 \leq \mu \leq 1$).

In this way, all the alternatives can be considered as theoretically available. If alternative j is not available ($\mu_I^i(j)=0$), the factor ($ln\mu_I^i(j)$) is such that its perceived

utility U^i_j tends to minus infinity and the probability of choosing it tends to zero, regardless of the value of V^i_j. Furthermore, choice probabilities of all the other alternatives are no longer influenced by alternative j. If, on the other hand, an alternative j is certainly available and taken into consideration ($\mu^i_I(j)=1$), the additional factor is equal to zero and the perceived utility has the conventional expression. Intermediate values of $\mu^i_I(j)$ reduce the utility of the alternative, proportionally to its level of availability.

The true value of the availability/perception level, and therefore of factor $ln\mu^i_I(j)$, for the generic individual i, is unknown to the analyst and can be simulated with a random variable. This in turn can be expressed by the sum of its mean value, $E[ln\mu^i_I(j)]$, and a random residual, η^i_j, given by the difference $ln\mu^i_I(j)$ - $E[ln\mu^i_I(j)]$. Expression (3.4.7) then becomes:

$$U^i_j = V^i_j + E[ln\mu^i_I(j)] + \eta^i_j + \varepsilon^i_j \tag{3.4.8}$$

In order to operationalize expression (3.4.8), the expected value of the logarithm of $\mu^i_I(j)$ can be approximately replaced by its second order Taylor series expression around the point $\bar{\mu}^i_I(j)=E[\mu^i_I(j)]$. Once the expectation is substituted in equation (3.4.8) it results:

$$U^i_j \cong V^i_j + ln\bar{\mu}^i_I(j) - \frac{1-\bar{\mu}^i_I(j)}{2\bar{\mu}^i_I(j)} + \sigma^i_j \qquad \text{with } \sigma^i_j = \varepsilon^i_j + \eta^i_j \tag{3.4.9}$$

The choice probability of alternative j can therefore be calculated by using the random utility models described in section 3.3 and will depend on the systematic utility of each alternative, on the mean availability/perception of each alternative and on the joint distribution of random variables σ^i_j. For example, if the latter are assumed to be i.i.d. Gumbel $(0,\theta)$ variables, a new Multinomial Logit model is obtained:

$$p^i[j] = \frac{exp\left[\frac{1}{\theta}\cdot\left(V^i_j + ln\bar{\mu}^i_I(j) - \frac{1-\bar{\mu}^i_I(j)}{2\bar{\mu}^i_I(j)}\right)\right]}{\sum_h exp\left[\frac{1}{\theta}\cdot\left(V^i_h + ln\bar{\mu}^i_I(h) - \frac{1-\bar{\mu}^i_I(h)}{2\bar{\mu}^i_I(h)}\right)\right]} \tag{3.4.10}$$

where the sum at the denominator is extended to all the alternatives theoretically available to decision-maker i. From the above expression, it can be deduced that the choice probability of a generic alternative increases as its mean availability/ perception increases[17] everything else being equal.

Other functional specifications of choice models can be obtained from expression (3.4.9). For example, a positive covariance between the residuals η_j and

η_h can be assumed if the two alternatives are more likely to be both available/ perceived or not.

To specify completely the model (3.4.10) or a different functional form, the mean availability/perception $\mu^i_j(j)$ must be expressed as a function of attributes of availability/perception, using for example a Binomial Logit model of type (3.4.6):

$$\mu^i_j(j) = \frac{1}{1 + exp\left(\sum_{k=1}^{K_j} \gamma_k Y^i_{kj}\right)} \tag{3.4.11}$$

Note the different interpretation of the two expressions (3.4.6) and (3.4.11). Expression (3.4.6) gives the probability that alternative j belongs to the choice set of a given decision-maker, while expression (3.4.11) gives the average degree of availability/perception of the alternative for decision-makers with the same attributes Y^i_{kj}.

3.5. Expected Maximum Perceived Utility and mathematical properties of random utility models

The Expected Maximum Perceived Utility (EMPU) is an important variable associated with each choice context. As was seen in section 3.2, random utility models are based on the assumption that perceived utilities are simulated as random variables and that the i^{th} decision-maker chooses alternative $j(i)$ with maximum perceived utility $U_{j(i)}$:

$$U_{j(i)} = max_j\{U^i_j\} = max(\boldsymbol{U}^i) \quad j \in I^i \tag{3.5.1}$$

The variable $U_{j(i)}$ therefore is the perceived utility "obtained" by the decision-maker in the choice context. This variable is not observed by the analyst because it is the maximum value of unobserved perceived utilities. Therefore $U_{j(i)}$ can be modeled as a random variable.

The Expected Maximum Perceived Utility (EMPU) associated with a given choice context is defined as the expected value of $U_{j(i)}$:

$$\begin{aligned} s^i = s^i(\boldsymbol{V}) &= E[U_{j(i)}] = E[max_j(\boldsymbol{U}^i)] = E[max(\boldsymbol{V}^i + \boldsymbol{\varepsilon}^i)] = \\ &= \int\ldots\int\ldots\int max(\boldsymbol{V}^i + \boldsymbol{\varepsilon}^i) f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} \end{aligned} \tag{3.5.2}$$

From (3.5.2) it can be deduced that EMPU is a function of the systematic utilities of all the alternatives, vector $\boldsymbol{V}^i$, and depends on the joint probability density function of the random residuals, $f(\boldsymbol{\varepsilon})$, as well as on the composition of choice set $I^{i(18)}$.

A number of mathematical properties of random utility model can be demonstrated using the EMPU variable. These properties are useful in building

transport demand model systems (see Chapter 4), for the analysis of assignment model (see Chapter 5) and for the evaluation of transport system projects (see Chapter 10).

In the following the two cases of probabilistic ($\varepsilon \neq 0$) and deterministic ($\varepsilon = 0$) choice models will be addressed separately.

Mathematical properties of probabilistic choice models. The EMPU associated with a particular choice context is always larger than, or equal to, the maximum systematic utility:

$$s(\boldsymbol{V}) \geq max(\boldsymbol{V}) \tag{3.5.3}$$

In fact, by definition, it results

$$s(\boldsymbol{V}) = \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} max(\boldsymbol{V}+\boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon}$$

and as $f(\boldsymbol{\varepsilon}) \geq 0$ and $max(\boldsymbol{V}+\boldsymbol{\varepsilon}) \geq V_k + \varepsilon_k \quad \forall k \in I$, it follows:

$$s(\boldsymbol{V}) = \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} max(\boldsymbol{V}+\boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} \geq$$

$$\geq \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} V_k\, f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} + \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} \varepsilon_k\, f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} =$$

$$= V_k \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} + \int_{\varepsilon_1=-\infty}^{\infty} \dots \int_{\varepsilon_m=-\infty}^{\infty} \varepsilon_k\, f(\boldsymbol{\varepsilon})\, d\boldsymbol{\varepsilon} =$$

$$= V_k + E[\varepsilon_k] = V_k \quad \forall k \in I$$

Therefore $s(\boldsymbol{V})$ is larger than, or equal to, the largest systematic utility, $s(\boldsymbol{V}) \geq V_k \; \forall k \in I$.

In addition, the mean systematic utility calculated as the mean of the systematic utilities of all alternatives k weighted with the respective choice probability $p_k(\boldsymbol{V})$ is less than or equal to the EMPU variable. In fact, using expression (3.5.3) it follows:

$$p(\boldsymbol{V})^T \boldsymbol{V} = \sum_k p_k(\boldsymbol{V}) V_k \leq \sum_k p_k(\boldsymbol{V}) max(\boldsymbol{V}) = max(\boldsymbol{V}) \leq s(\boldsymbol{V})$$

In order to analyze the EMPU variable in more detail, initially, reference can be made to a Multinomial Logit model with constant parameter θ. For this model $s(\boldsymbol{V})$ can be expressed in closed form. In fact, according to the results reported for the maximization of Gumbel variables[19], the EMPU is given by expression (3.5.5) repeated here:

$$s(\boldsymbol{V}) = \theta\, ln \sum_j exp(V_j/\theta) \tag{3.5.4}$$

It can easily be deduced that expression (3.5.4) satisfies condition (3.5.3) as exemplified in Fig. 3.5.1. From expression (3.5.4) it is also deduced that the EMPU for a Multinomial Logit model increases if the systematic utility of one or more alternatives increases since the functions $ln(.)$ and $exp(.)$ are both monotonic increasing. Furthermore, because of the non-negativity of the exponential function,

EMPU increases with the number of available alternatives. The addition of a new alternative to the choice set, results in an increase in the expected maximum perceived utility, even if the new alternative has a systematic utility inferior to those already available. In fact, because of the randomness of perceived utilities, some decision-maker will perceive the new alternative as the alternative of maximum utility, for these individuals the $max_j(U^i)$ clearly increases with a consequent general increase of the mean value over all the individuals of $max_j(U^i)$, that is the EMPU variable. The example in Fig. 3.5.1 exemplifies also this point.

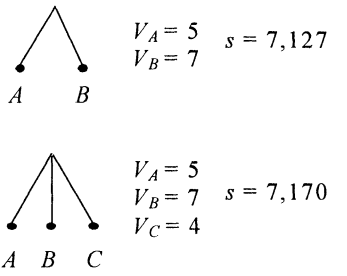


Fig. 3.5.1 Example of calculation of the Expected Maximum Perceived Utility (EMPU).

These properties of EMPU, directly derived for the Multinomial Logit, apply also to the largest class of *additive random utility models*. For these models, the density function of random residuals doesn't depend on V:

$$f(\varepsilon/V) = f(\varepsilon) \qquad \forall \varepsilon \in E^m \tag{3.5.5}$$

All the random utility models described in section 3.3 are additive if the parameters of the $f(\varepsilon)$, do not depend on the vector V. If the joint density function of the random residuals $f(\varepsilon)$ is continuous with its first derivatives, choice probabilities $p(V)$, and the EMPU, $s(V)$, are also continuous functions of V, together with their first derivatives. All random utility models described in section 3.3 satisfy these continuity requirements. Under these assumptions, additive random utility models share some general mathematical properties connected with the Expected Maximum Perceived Utility.

1) The *partial derivative* of the EMPU variable with respect to the systematic utility V_k is equal to the choice probability of alternative k:

$$\frac{\partial s(V)}{\partial V_k} = p[k](V) \tag{3.5.6}$$

The gradient of EMPU is thus equal to the vector of choice probabilities:

$$\nabla s(V) = p(V) \tag{3.5.7}$$

and its Hessian is equal to the Jacobian of choice probabilities:

$$\boldsymbol{Hess}[s(\boldsymbol{V})] = \boldsymbol{Jac}[p(\boldsymbol{V})] \tag{3.5.7a}$$

In fact, for a continuous function with continuous first derivatives, the integration and differentiation operators can be exchanged:

$$\begin{aligned}\frac{\partial s(\boldsymbol{V})}{\partial V_k} &= \frac{\partial}{\partial V_k}\int_{\varepsilon_1=-\infty}^{\infty}\dots\int_{\varepsilon_m=-\infty}^{\infty} max(\boldsymbol{V}+\boldsymbol{\varepsilon})f(\boldsymbol{\varepsilon})d\boldsymbol{\varepsilon} = \\ &= \int_{\varepsilon_1=-\infty}^{\infty}\dots\int_{\varepsilon_m=-\infty}^{\infty}\frac{\partial max(\boldsymbol{V}+\boldsymbol{\varepsilon})}{\partial V_k}f(\boldsymbol{\varepsilon})d\boldsymbol{\varepsilon}\end{aligned} \tag{3.5.8}$$

Since it results:

$$\frac{\partial max(\boldsymbol{V}+\boldsymbol{\varepsilon})}{\partial V_k} = \begin{cases}1 \text{ for } k \text{ such that } V_k+\varepsilon_k = max(\boldsymbol{V}+\boldsymbol{\varepsilon}) \\ 0 \text{ otherwise}\end{cases}$$

the integral (3.5.8) is equal to the probability that the perceived utility of alternative k, $V_k + \varepsilon_k$, is the largest among all the m alternatives available, from which expression (3.5.6) derives.

This result can be checked immediately for the Multinomial Logit model whose EMPU, expressed by (3.5.4), can be differentiated analytically:

$$\frac{\partial}{\partial V_k}\left[\theta ln\sum_j exp(V_j/\theta)\right] = \frac{exp(V_k/\theta)}{\sum_j exp(V_j/\theta)} = p[k](\boldsymbol{V}) \tag{3.5.9}$$

Furthermore, since the choice probability $p[k]$ is always greater than or equal to zero, according to (3.5.6) the derivative of EMPU with respect to the systematic utility is always non-negative, EMPU increases (or does not decrease) as the systematic utility of each alternative increases and, by extension, increases as the number of available alternatives increases[20].

2) The EMPU function is *convex*[21] with respect to the vector of systematic utilities $\boldsymbol{V}$.

In fact, for each $\boldsymbol{\varepsilon}$, $f(\boldsymbol{\varepsilon}) \geq 0$ and $\max(\boldsymbol{V} + \boldsymbol{\varepsilon})$ is a convex function of $\boldsymbol{V}$; it follows that the Expected Maximum Perceived Utility function $s(\boldsymbol{V})$ expressed by (3.5.2) is a linear combination with non-negative coefficients of convex functions, and therefore it is convex too.

Note that in virtue of property 2) the EMPU function has a Hessian matrix, $\boldsymbol{Hess}(s(\boldsymbol{V}))$, which is (symmetric and) positive semidefinite. Consequently, the Jacobian of choice probabilities, $\boldsymbol{Jac}(\boldsymbol{p}(\boldsymbol{V}))$, is (symmetric and) positive semidefinite (see equation 3.5.7a).

3) If the EMPU function is continuous and differentiable it results:

$$s(\boldsymbol{V}') \geq s(\boldsymbol{V}'') + p(\boldsymbol{V}'')^T(\boldsymbol{V}'-\boldsymbol{V}'') \quad \forall \boldsymbol{V}', \boldsymbol{V}'' \tag{3.5.10a}$$

and choice probabilities are monotonic increasing functions of systematic utilities.

$$(p(\boldsymbol{V}') - p(\boldsymbol{V}''))^T(\boldsymbol{V}'-\boldsymbol{V}'') \geq 0 \quad \forall \boldsymbol{V}', \boldsymbol{V}'' \qquad (3.5.10b)$$

In fact, because the EMPU function is convex and differentiable, it follows that:

$$s(\boldsymbol{V}') \geq s(\boldsymbol{V}'') + \nabla s(\boldsymbol{V}'')^T(\boldsymbol{V}'-\boldsymbol{V}'') \quad \forall \boldsymbol{V}', \boldsymbol{V}''$$

and its gradient must be an increasing monotonic function (see Appendix A):

$$(\nabla s(\boldsymbol{V}') - \nabla s(\boldsymbol{V}''))^T(\boldsymbol{V}'-\boldsymbol{V}'') \geq 0 \quad \forall \boldsymbol{V}', \boldsymbol{V}''$$

Applying (3.5.7), the two preceding expressions can be formulated in terms of the vector of choice probabilities as in (3.5.10a). Moreover from (3.5.10a) it results:

$$s(\boldsymbol{V}') - s(\boldsymbol{V}'') \geq p(\boldsymbol{V}'')^T(\boldsymbol{V}'-\boldsymbol{V}'') \quad \forall \boldsymbol{V}', \boldsymbol{V}''$$

$$s(\boldsymbol{V}'') - s(\boldsymbol{V}') \geq p(\boldsymbol{V}')^T(\boldsymbol{V}''-\boldsymbol{V}') \quad \forall \boldsymbol{V}', \boldsymbol{V}''$$

Summing terms by terms last two equations it yields:

$$0 \geq p(\boldsymbol{V}'')^T(\boldsymbol{V}'-\boldsymbol{V}'') + p(\boldsymbol{V}')^T(\boldsymbol{V}''-\boldsymbol{V}') \quad \forall \boldsymbol{V}', \boldsymbol{V}''$$

from which the (3.5.10b) is easily obtained.

In particular, equation (3.5.10b) can be expressed for a single alternative, assuming that the systematic utilities of all other choice alternatives are constant:

$$p_k(V_k') \geq p_k(V_k'') \quad \text{if} \quad V_k' \geq V_k''$$

In other words, the choice probability of a generic alternative does not decrease as its systematic utility increases, if all the other systematic utilities remain unchanged. Using an analogous argument, it can be demonstrated that as V_k tends to minus infinity, the choice probability of alternative k tends to zero:

$$\lim_{V_k \to -\infty} p[k] = 0$$

Mathematical properties of the deterministic choice model. The *deterministic choice model*[22] is obtained if the random residuals are all equal to zero. In this case the perceived utility coincides with the systematic utility and only the alternative(s) of maximum utility can be chosen:

$$p[k] > 0 \Rightarrow V_k = max(\boldsymbol{V})$$

and

$$V_k = max(\boldsymbol{V}) \Rightarrow p[k] \in [0,1], \quad V_k < max(\boldsymbol{V}) \Rightarrow p[k] = 0$$

Note that the deterministic choice model satisfies condition (3.5.5) and can therefore be considered an additive model. If there are two or more alternatives with

(equal) maximum systematic utility, there are many vectors of choice probabilities satisfying the above conditions. In this case, the relation $\boldsymbol{p}(\boldsymbol{V})$ is not a function, but a one-to-many map. Let $\boldsymbol{p}_{DET}(\boldsymbol{V})$ be one of the possible vectors of choice probabilities corresponding to vector $\boldsymbol{V}$ through the deterministic choice map.

The following necessary and sufficient condition guarantees that a probability vector, $\boldsymbol{p}^*$ with $\boldsymbol{p}^* \geq \boldsymbol{0}$, $\boldsymbol{1}^T\boldsymbol{p}^* = 1$, is a deterministic choice probability vector:

$$\boldsymbol{p}^* = \boldsymbol{p}_{DET}(\boldsymbol{V}) \Leftrightarrow \boldsymbol{V}^T\boldsymbol{p}^* = max(\boldsymbol{V})\ \boldsymbol{1}^T\boldsymbol{p}^* = max(\boldsymbol{V}) \qquad (3.5.11a)$$

In fact, given a vector of deterministic probabilities, $\boldsymbol{p}^* = \boldsymbol{p}_{DET}(\boldsymbol{V})$, it follows that $\boldsymbol{V}^T\boldsymbol{p}^* = max(\boldsymbol{V})$ as p_k^* can be positive only for an alternative k of maximum systematic utility, and vice-versa. Furthermore, condition $\boldsymbol{1}^T\boldsymbol{p}^* = 1$ implies that $max(\boldsymbol{V})\boldsymbol{1}^T\boldsymbol{p}^* = max(\boldsymbol{V})$.

In general, for any vector of choice probabilities, $\boldsymbol{p}$, since $\boldsymbol{1}^T\boldsymbol{p} = 1$, then, as observed earlier:

$$\boldsymbol{V}^T\boldsymbol{p} \leq max(\boldsymbol{V})\ \boldsymbol{1}^T\boldsymbol{p} = max(\boldsymbol{V}) \qquad \forall \boldsymbol{p}: \boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{1}^T\boldsymbol{p} = 1$$

Consistently with (3.5.11a), in the above relationship the sign of equality holds only for a vector of deterministic probabilities. Combining the above relationship with (3.5.11a), a basic relationship for deterministic assignment models described in Chapter 5 can be obtained:

$$(\boldsymbol{V} - max(\boldsymbol{V})\boldsymbol{1})^T(\boldsymbol{p} - \boldsymbol{p}_{DET}(\boldsymbol{V})) \leq 0 \qquad \forall \boldsymbol{p}: \boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{1}^T\boldsymbol{p} = 1 \qquad (3.5.11b)$$

The deterministic utility model has the properties 2) and 3) described above for probabilistic and additive models[23]. In particular, for what concern property 2), the Expected Maximum Perceived Utility for a deterministic model is a convex function of systematic utilities and is equal to the maximum systematic utility:

$$s(\boldsymbol{V}) = max(\boldsymbol{V}) = \boldsymbol{p}_{DET}(\boldsymbol{V})^T\boldsymbol{V} \qquad (3.5.12)$$

This condition and result (3.5.3) imply that EMPU for a deterministic choice model is less than or equal to the EMPU for any probabilistic choice model for a given vector of systematic utilities $\boldsymbol{V}$. A behavioral interpretation of this result suggests that the presence of random residuals makes the perceived utility for the chosen alternative, on average, larger than the systematic utility of this alternative which is the perceived utility in a deterministic choice model.

For what concern property 3), the deterministic choice map is non-decreasing monotonic with respect to systematic utilities, as for additive probabilistic choice functions:

$$s(\boldsymbol{V}') \geq s(\boldsymbol{V}'') + p_{DET}(\boldsymbol{V}'')^T(\boldsymbol{V}' - \boldsymbol{V}'') \qquad \forall \boldsymbol{V}', \boldsymbol{V}'' \qquad (3.5.13a)$$

or:

$$(p_{DET}(V') - p_{DET}(V''))^T (V'-V'') \geq 0 \quad \forall V', V'' \qquad (3.5.13b)$$

in perfect formal analogy with expressions (3.5.10).

In fact, from (3.5.11a) it follows:

$$max(V') = (V')^T \boldsymbol{p}_{DET}(V')$$
$$max(V'') = (V)^T \boldsymbol{p}_{DET}(V'')$$

Subtracting term by term last two equations, it results:

$$max(V') - max(V'') = (V')^T \boldsymbol{p}_{DET}(V') - (V'')^T \boldsymbol{p}_{DET}(V'') \qquad (i)$$

Since:

$$(V')^T \boldsymbol{p}_{DET}(V') = max(V') \geq (V')^T \boldsymbol{p} \quad \forall \boldsymbol{p},$$

for $\boldsymbol{p} = \boldsymbol{p}_{DET}(V'')$ it follows:

$$(V')^T \boldsymbol{p}_{DET}(V') \geq (V')^T \boldsymbol{p}_{DET}(V'')$$

from which:

$$(V')^T \boldsymbol{p}_{DET}(V') - (V'')^T \boldsymbol{p}_{DET}(V'') \geq (V')^T \boldsymbol{p}_{DET}(V'') - (V'')^T \boldsymbol{p}_{DET}(V'') \qquad (ii)$$

Therefore, combining equations (i) and (ii), it yields

$$max(V') - max(V'') \geq (V'-V'')^T \boldsymbol{p}_{DET}(V'')$$

which is expression (3.5.13a) since $s(V) = max(V)$.

3.6. Direct and cross elasticities of random utility model

In every respect, random utility models can be considered demand functions in the econometric sense. In fact, choice probabilities can be seen as the mean values of the fractions of a certain market segment (a group of decision-makers with the same characteristics) using each alternative[24]. Also, these fractions are expressed as function of the attributes of the available alternatives. In the context of this interpretation, it is possible to extend to random utility models the microeconomic concepts of direct and cross elasticities of demand functions with respect to infinitesimal or discrete variations of the variables in the utility function.

Recall that *direct elasticity* is defined as the percentage variation of the demand for a certain commodity (in this case, of the choice probability of alternative j) divided by the percentage variation of a variable (attribute) relative to the same commodity X_{kj}:

$$E_{kj}^{p[j]} = \frac{\Delta p[j]}{p[j]} \Big/ \frac{\Delta X_{kj}}{X_{kj}}$$

Analogously, *cross elasticity* is defined as the percentage variation of the demand for a certain commodity j divided by the percentage variation of a variable k relative to another commodity h, X_{kh}:

$$E_{kh}^{p[j]} = \frac{\Delta p[j]}{p[j]} \Big/ \frac{\Delta X_{kh}}{X_{kh}}$$

In the above definitions, the variations of attributes and demand are assumed to be finite. In this case, we speak of arc elasticity, which is calculated as the ratio of incremental ratios over an "arc" of the demand curve. Point elasticities are defined for infinitesimal variations and can be expressed analytically.

The point direct elasticity of the choice probability for alternative j with respect to an infinitesimal variation of the k^{th} attribute of the utility function of this alternative, X_{kj}, is defined as:

$$E_{kj}^{p[j]} = \frac{\partial p[j](\boldsymbol{X})}{\partial X_{kj}} \frac{X_{kj}}{p[j]} = \frac{\partial lnp[j](\boldsymbol{X})}{\partial lnX_{kj}} \tag{3.6.1}$$

where $\boldsymbol{X}$ includes the vectors of attributes for all alternatives.

Similarly the point cross elasticity of the choice probability of alternative j with respect to an infinitesimal variation of the kth attribute of the utility function of the alternative h, X_{kh}, can be defined as:

$$E_{kh}^{p[j]} = \frac{\partial p[j](\boldsymbol{X})}{\partial X_{kh}} \frac{X_{kh}}{p[j]} = \frac{\partial lnp[j](\boldsymbol{X})}{\partial lnX_{kh}} \tag{3.6.2}$$

Both direct and cross elasticities[25] are useful measures of the model's sensitivity to variations of the attributes. It is evident from (3.6.1) and (3.6.2) that elasticities depend on the functional form of the model as well as on the values of attributes and parameters in the systematic utilities.

Analytic and compact expressions of direct and cross elasticities (3.6.1) and (3.6.2) can be obtained for the Multinomial Logit model with linear systematic utility function $V_j = \beta^T \boldsymbol{X}_j$. In this case, it results:

$$E_{kj}^{p[j]} = (1 - p[j])\beta_k X_{kj} / \theta \tag{3.6.3}$$

$$E_{kh}^{p[j]} = -p[k]\beta_k X_{kh} / \theta \tag{3.6.4}$$

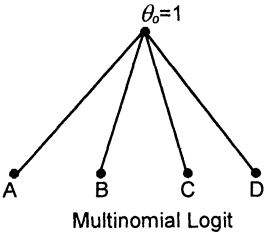
From (3.6.3) it can be deduced that the direct elasticity is positive if attribute X_{kj} is positive (as it is usually the case) and if its coefficient β_k is positive. In other words, the choice probability of an alternative increases if the value of an attribute representing an utility (β positive) increases[26]. The increase will be higher the higher the values of coefficient β_k and attribute X_{kj} and the lower the value of the choice probability of alternative j. Thus in a mode choice model, direct elasticities of the probability of choosing the car in terms of travel time and cost will be negative since the coefficients β_k of these attributes are negative; these elasticities will be larger in absolute terms, for an Origin-Destination pair with relatively large time and cost values. Lastly, if the probability of choosing the car is low, its elasticity will be larger, for given values of parameter β_k and attribute X_{kj}.

Similar considerations, though with inverted signs, hold for cross elasticities, which will be positive if β_k or X_{kh} are negative and larger the larger in absolute value are β_k, X_k and $p[h]$. Continuing with the above example, cross elasticities of the probability of using the car with respect to travel time and cost of another mode will be positive ($\beta_k<0$).

Qualitatively similar considerations apply to elasticities for random utility models other than *MNL*.

Note that the cross elasticity (3.6.4) of the Multinomial Logit model is constant for all alternatives j as the variation of an attribute of a certain alternative h produces the same percentage variation in the choice probabilities of all other alternatives. This result can be considered as a different manifestation of the property of Independence from Irrelevant Alternatives of the Logit model described in section 3.3.1.

Expression (3.6.3) and (3.6.4) also show that, for given values of coefficients and attributes, direct and cross elasticities are higher in absolute terms when the variance of random residuals (parameter θ) is lower. Vice versa, for variances tending to infinity, elasticities tend to zero. Fig. 3.6.1 shows the values of direct and cross elasticities with respect to a generic attribute in a Multinomial Logit model.



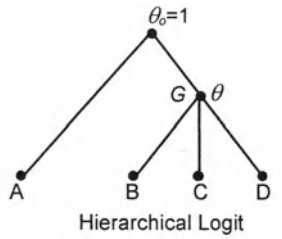
$V_A = V_B = V_C = V_D$
$\beta_k = 1 \quad \forall k$
$X_{kj} = 1 \quad \forall k, j$

	X_{kA}	X_{kB}	X_{kC}	X_{kD}
$E_{p[A]}$	0.75	-0.25	-0.25	-0.25
$E_{p[B]}$	-0.25	0.75	-0.25	-0.25
$E_{p[C]}$	-0.25	-0.25	0.75	-0.25
$E_{p[D]}$	-0.25	-0.25	-0.25	0.75

Fig. 3.6.1 Direct and cross elasticities for a Multinomial Logit model.

For more complex random utility models it is not easy, or even possible, to derive analytic expressions of direct and cross elasticities. However, it is useful to discuss elasticities for a Single-Level Hierarchical Logit model since they provide some insight into the influence of covariances on direct and cross elasticities.

For the Single-Level Hierarchical Logit model in Fig. 3.6.2 it is possible to express the elasticities of alternative A, the only component of a group, with respect to the variation of a generic attribute X_k included in the systematic utility of all alternatives.



Hierarchical Logit

	θ=1				θ=0.8				θ=0.4				θ=0.1			
	X_{kA}	X_{kB}	X_{kC}	X_{kD}	X_{kA}	X_{kB}	X_{kC}	X_{kD}	X_{kA}	X_{kB}	X_{kC}	X_{kD}	X_{kA}	X_{kB}	X_{kC}	X_{kD}
$E_{p[A]}$	0.75	-0.25	-0.25	-0.25	0.71	-0.24	-0.24	-0.24	0.61	-0.20	-0.20	-0.20	0.52	-0.18	-0.18	-0.18
$E_{p[B]}$	-0.25	0.75	-0.25	-0.25	-0.29	0.93	-0.32	-0.32	-0.39	1.80	-0.70	-0.70	-0.47	6.82	-3.18	-3.18
$E_{p[C]}$	-0.25	-0.25	0.75	-0.25	-0.29	-0.32	0.93	-0.32	-0.39	-0.70	1.80	-0.70	-0.47	-3.18	6.82	-3.18
$E_{p[D]}$	-0.25	-0.25	-0.25	0.75	-0.29	-0.32	-0.32	0.93	-0.39	-0.70	-0.70	1.80	-0.47	-3.18	-3.18	6.82

Fig. 3.6.2 Direct and cross elasticities for a Hierarchical Logit model.

Applying the definitions of elasticity (3.6.1) and (3.6.2) to the Single-Level Hierarchical Logit model in expression (3.3.19) with parameter θ_o=1, it is possible to obtain the elasticities of alternative A with respect to attribute k in closed form. The direct elasticity (variation of attribute X_{kA}) and the cross elasticity with respect to alternative B belonging to group G, also including alternatives C and D (variation of attribute X_{kB}) are respectively:

$$E_{kA}^{p[A]} = (1 - p[A])\beta_k X_{kA} / \theta \tag{3.6.5}$$

$$E_{kB}^{p[A]} = -p[B]\beta_k X_{kB} / \theta \tag{3.6.6}$$

The elasticities in this case are completely analogous to those obtained for the Multinomial Logit model, expressed by equations (3.6.3) and (3.6.4). Things are different for the direct elasticity X_{kB} of alternative B belonging to group G:

$$E_{kB}^{p[B]} = \{(1 - p[G]) \cdot p[B/G] + (1 - p[B/G])/\theta\}\beta_k X_{kB} \tag{3.6.7}$$

If the Hierarchical Logit reduced to a Multinomial Logit model, i.e. if θ = 1, direct elasticity (3.6.7) would become analogous to (3.6.3) or (3.6.5). On the other hand, if θ is less than one, the Hierarchical Logit elasticity is larger than that obtained for a Multinomial Logit with the same parameters, attributes and residuals variance.

The cross elasticities of $p[B]$ with respect to variations of the attribute X_{kA} of the "isolated" alternative A, and X_{kC} of the alternative C belonging to the group G are respectively:

$$E_{X_{kA}}^{p[B]} = -p[A]\beta_k X_{kA} \tag{3.6.8}$$

$$E_{X_{kC}}^{p[B]} = -\left[p[C] + \frac{1-\theta}{\theta} p[C/G] \right] \beta_k X_{kC} \qquad (3.6.9)$$

Equation (3.6.8) shows that the cross elasticity with respect to an attribute of alternative *A not* belonging to group *G*, is equivalent to that of the corresponding Multinomial Logit model. On the other hand, cross elasticity with respect to an attribute of an alternative belonging to group *G* (correlated with *B*) is larger for smaller values of parameter θ, i.e. the larger the covariance between the two alternatives. If two alternatives are perceived as being very similar (i.e. their respective random residuals are highly correlated), the probability of choosing one of them is very sensitive to variations of the attributes of the other. From (3.6.9) it also results that if $\theta = 1$, the Hierarchical Logit model becomes a Multinomial Logit model and the cross elasticity is analogous to (3.6.8).

Direct and cross elasticities of the Hierarchical Logit model for different values of parameter θ, are shown in Fig. 3.6.2. For $\theta = 1$ the elasticities reported in Fig. 3.6.1 are obtained.

The general deductions from the above example are that, given equal attributes and coefficients, direct and cross elasticities of an alternative are the higher the more that alternative is perceived as "similar" to other alternatives. Thus for any random utility model, the variation of an attribute of an alternative will produce more sensible variations in the probability of choosing alternatives perceived as closer substitutes.

3.7. Aggregation methods for random utility models

Random utility models described in previous sections express the probability that a decision-maker *i* chooses each alternative *j*, as a function of the attributes of all available alternatives. To highlight the dependence of choice probabilities on individual attributes, expression (3.2.3a) can be reformulated as:

$$p^i[j/V(\boldsymbol{X}^i)] = Pr[V_j(\boldsymbol{X}^i_j) + \varepsilon^i_j \geq V_k(\boldsymbol{X}^i_k) + \varepsilon^i_k \quad \forall k \in I^i] \qquad (3.7.1)$$

where $\boldsymbol{X}^i_j$ is the vector of attributes of alternative *j* for decision-maker *i*, and $\boldsymbol{X}^i$ the vector of the attributes of all alternatives. For convenience of notation, from now on (3.7.1) will be represented more compactly as $p[j/\boldsymbol{X}^i]$.

Applications of random utility models to the simulation of travel demand often require the mean value of total demand flows, i.e. the mean number of decision-makers choosing each alternative. Aggregation techniques allow to pass from individual choice probabilities to group, or aggregate probabilities. To introduce these techniques, it is useful to describe the theoretical aggregation process. Suppose that, for each individual *i* of the population, the vector $\boldsymbol{X}^i$ of attributes, the functional form and the coefficients of the random utility model are known. Suppose also that there are N_T individuals in the population and that they choose independently. Under these assumptions, the number of decision-makers who actually choose the generic

alternative j is a random variable, the sum of N_T independent Bernoulli random variables y_j^i, each of which is equal to one if individual i chooses alternative j, zero otherwise. The mean value of the number of individuals choosing alternative j, D_j, is therefore the sum of the means, $p[j/\boldsymbol{X}^i]$, of the N_T Bernoulli random variables:

$$D_j = \sum_{i=1}^{N_T} E[y_j^i] = \sum_{i=1}^{N_T} p[j / \boldsymbol{X}^i] \tag{3.7.2}$$

The fraction, or the average percentage, of the population choosing alternative j, P_j, can be estimated as:

$$P_j = \frac{1}{N_T} \sum_{i=1}^{N_T} p[j / \boldsymbol{X}^i] = \frac{D_j}{N_T} \tag{3.7.3}$$

For populations large enough to replace the sum with the integral, equation (3.7.3) can be rewritten as:

$$P_j = \int_{\boldsymbol{X}} p[j / \boldsymbol{X}] g(\boldsymbol{X}) d\boldsymbol{X} \tag{3.7.4}$$

where $g(\boldsymbol{X})$ represents the joint probability density function of the vector of attributes over the whole population, a measure of the frequency with which the different values of $\boldsymbol{X}$ occur in the population. In practice, the distribution $g(\boldsymbol{X})$ is not known, and to calculate the percentage P_j, aggregation techniques allowing to find an estimate $\hat{P}_j$ using information on a limited number of individuals must be used.

In the literature, various aggregation methods have been proposed; these can be seen as approximate integration techniques of equation (3.7.4).

The methods most frequently applied are:

1) average individual;
2) classification;
3) sample enumeration;
4) classification/enumeration.

1) In the first method, an "*average individual*" is considered, whose attributes $\overline{\boldsymbol{X}}$ are the average values over the population with respect to the $g(\boldsymbol{X})$. The aggregated choice percentage is calculated as a function of these attributes:

$$\hat{P}_j = p[j / \overline{\boldsymbol{X}}] \tag{3.7.5}$$

This method is acceptable only if the relationship between the vector of attributes and the choice probabilities, $p[j/\boldsymbol{X}]$, is linear or almost linear. Should the

probability function be convex or concave, the method would respectively underestimate or overestimate the actual value of the fraction of the population choosing alternative j (see Fig. 3.7.1). It can also be shown that the deviation of linear estimate $\hat{P}_j$ from its true value is the larger the more dispersed are the values of X in the population, i.e. the larger the variances in the marginal distributions of $g(\boldsymbol{X})$.

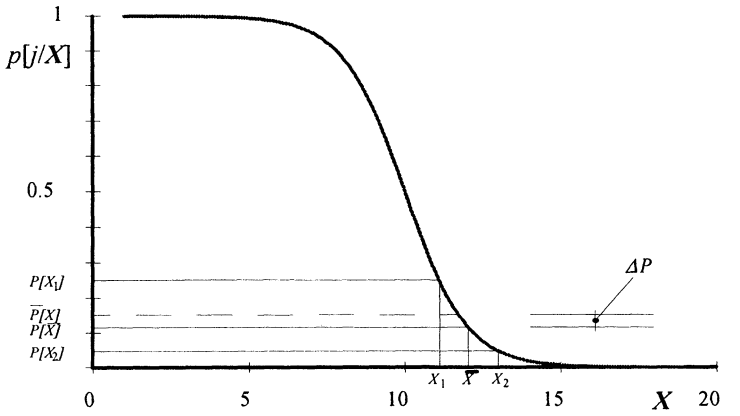


Fig. 3.7.1 Bias of average individual estimates of population fractions.

2) The *classification* method can be seen as an extension of the average individual method described above. In order to reduce the variance of $g(\boldsymbol{X})$, the population is divided into homogeneous and mutually exclusive classes, with i representing the generic class of N_i components; the average individual technique is applied to each class. The estimated value of the fraction of population choosing alternative j therefore becomes:

$$\hat{P}_j = \sum_{i=1}^{I} \frac{N_i}{N_T} P[j / \bar{\boldsymbol{X}}^i] \tag{3.7.6}$$

where $\bar{\boldsymbol{X}}^i$ is the vector of attributes for the average individual of the i^{th} class.

In applications, classes are defined on the basis of few main criteria having the greatest effect on systematic utilities. Variables influencing the distribution of the attributes, are often adopted as classification criteria, e.g. professional status or income. The number N_i of individuals belonging to each class should be available from statistical sources. The classification technique gives satisfactory results when the number of classes is limited and classes show good internal homogeneity with respect to the attributes included in the model.

3) With the *sample enumeration* method, it is assumed that the whole population can be represented by a random sample of individuals (decision-makers) extracted

from it. The average fraction of individuals choosing alternative j in the whole population, is estimated starting from the probability that j is chosen by the individuals belonging to the random sample. If N_s is the number of individuals in the sample, then:

$$\hat{P}_j = \frac{1}{N_s} \sum_{h=1}^{N_s} p[j / \boldsymbol{X}^h] \tag{3.7.7}$$

where $\boldsymbol{X}^h$ is the vector of the attributes relative to the h^{th} individual in the sample. Expression (3.7.7) is relative to the estimation of the mean percentage of the population in the case of simple random sampling [27].

4) *Sample enumeration and classification* methods can be combined; this is equivalent to assuming a stratified random sample of decision-makers. A random sample of individuals is extracted from each of the I strata in which the population is divided. If N_i is the number of individuals belonging to stratum i and N_{si} is the number of sample individuals extracted from stratum i, the faction $\hat{P}_j$ can be estimated as:

$$\hat{P}_j = \sum_{i=1}^{I} W_i \frac{1}{N_{si}} \sum_{h=1}^{N_{si}} p[j / \boldsymbol{X}^h] \tag{3.7.8}$$

where the ratio $W_i=N_i/N_T$ is the weight of stratum i in the population. The total number of decision-makers choosing each alternative j (aggregate demand for alternative j) can be calculated by multiplying expressions (3.7.6), (3.7.7) and (3.7.8) by N_T. The ratio between the number of individuals in the population (or a class) and the number of individuals in the sample, N_T/N_s or N_i/N_{si}, is called "expansion factor" of individuals from the sample to the population.

The sample enumeration method allows significant flexibility in the use of random utility models, since the attributes considered in vector $\boldsymbol{X}$ might include variables relating to the individual for which it is difficult, if not impossible, to obtain mean values over the whole population or sub-populations (classes). This flexibility is achieved at the cost of greater computational complexity. However, this drawback is becoming less important with the steady increase in available computing power. Another problem related to the sample enumeration method involves the availability of samples of decision-makers for each class i and each choice context (e.g., each traffic zone in the study area). The samples should be large enough to guarantee sufficient coverage of the distribution of attributes $\boldsymbol{X}$. This leads to the need for large samples of decision-makers for each zone. To overcome this problem, the *prototypical sample* method has been proposed. The same sample of N_{si} decision-makers of class i is used for different traffic zones while different weights W^z_i are adopted for each class i in each zone z ($W^z_i = N^z_i/N_T$). This method is based on knowledge of the number, N^z_i, of individuals of class i in each zone which

can be obtained from statistical sources (present scenario), or from socio-demographic forecasts (future scenarios).

Estimation of the average number of individuals choosing alternative j in zone z, D^z_j, with the sample enumeration method requires the expansion factors g^z_i of each category in each zone:

$$D^z_j = \sum_{i=1}^{I} g^z_i \sum_{h=1}^{N_{si}} p[j / \boldsymbol{X}^h] \tag{3.7.9}$$

where the expansion factors can be formally expressed as:

$$g^z_i = \frac{N^z_i}{N_{si}}$$

Sometimes, the number N^z_i of individuals of class i in zone z is unknown, especially when several classes have been defined. In this case, it is not possible to estimate the weights of the individual classes ($W^z_i = N^z_i/N_T$) and the average choice percentages by (3.7.8). Additionally, the expansion factors g^z_i and the total number of individuals choosing alternative j, D_j, cannot be estimated by (3.7.9). To overcome this drawback, the *target variable method* can be adopted. This method will be described in reference to the calculation of expansion factors; once these are known, the weights W^z_i can easily be calculated. The expansion factors are calculated so that once the prototypical sample is rescaled to its universe it reproduces the values of some aggregated variables, known as target variables, T^z_t. Typical target variables are the number of residents by professional status, age, sex, income group, etc. Formally expansion factors g^z_i must satisfy the following equations:

$$\sum_i g^z_i \sum_{h=1}^{N_{si}} K(t,h) = T^z_t \tag{3.7.10}$$

where $K(t, h)$ is the contribution to the t^{th} target variable of the h^{th} component of the prototypical sample belonging to category i. For example, if the t^{th} target variable is the number of workers in the zone, individual h of category i will contribute one if employed, zero otherwise. In general, the number of unknown expansion factors, or classes in each zone, is larger than the number N_t of target variables, and the system of equations (3.7.10) does not have a unique solution. In this case, the vector of expansion factors for each category $\boldsymbol{g}^z$ can be obtained by solving a Least Square problem minimizing the weighted distance from a vector of reference expansion factors $\hat{\boldsymbol{g}}$ and, at the same time, satisfying as closely as possible the system of equations (3.7.10):

$$\boldsymbol{g}^z = \underset{\boldsymbol{g}^z \geq 0}{argmin}\left[\sum_i (g_i^z - \hat{g}_i)^2 + \alpha \sum_{t=1}^{N_t}\left(\sum_i g_i^z \sum_{h=1}^{N_{si}} K(t,h) - T_t^z\right)^2\right] \qquad (3.7.11)$$

Reference expansion factors can be obtained as sample estimates of the users' fraction belonging to each category. The parameter α is the relative weight of the two parts of the objective function in (3.7.11), i.e. the relative weight that the analyst associates to the target variables (3.7.10) and to the initial estimates $\hat{\boldsymbol{g}}$ in the solution of problem (3.7.11).

The Least Square problem with non-negativity constraints on the variables (3.7.11) is similar to that formulated for the estimation of O-D demand flows with traffic counts discussed in Chapter 8, and can be solved by using the projected gradient algorithm described in Appendix A.

3.A. Derivation of Logit models from the GEV model

As stated in section 3.3.5 the choice probability of a GEV model can be expressed as (see equation 3.3.52):

$$p[j] = \frac{e^{V_j} \cdot G_j(e^{V_1},...,e^{V_j},...,e^{V_m})}{\mu \cdot G(e^{V_1},...,e^{V_j},...,e^{V_m})} \tag{3.A.1}$$

In the same section it was also stated that Multinomial Logit, Hierarchical Logit and Cross-Nested Logit models can be derived as GEV models. For the Multinomial Logit and the Hierarchical Logit this is possible by specifying the function $G(.)$ as:

$$G(e^{V_1},...,e^{V_m}) = e^{Y_o} \tag{3.A.2}$$

where Y_o, is the logsum variable relative to the root node of the choice tree related to the model under study.

3.A.1. Derivation of the Multinomial Logit model

In the case of the *Multinomial Logit* model, the choice tree has the root node o directly connected to all the elementary alternatives j (see Fig. 3.3.1).

In this case the variable Y_o can be expressed as:

$$Y_o = ln\sum_{i=1}^{m} e^{V_i/\theta}$$

and equation (3.A.7) becomes:

$$G(e^{V_1},...,e^{V_m}) = \sum_{i=1}^{m} e^{V_i/\theta} \tag{3.A.3}$$

It can easily be verified that this function satisfies the four properties mentioned in section 3.3.5, given some limitations to parameter θ.

In fact:

1) $G \geq 0$ for any value of θ and V_i ($i = 1,...,m$);

2) $G(\alpha e^{V_1},...,\alpha e^{V_m}) = \sum_{i=1}^{m} (\alpha e^{V_i})^{1/\theta} = \alpha^{1/\theta} \sum_{i=1}^{m} (e^{V_i})^{1/\theta} = \alpha^{1/\theta} G(e^{V_1},...,e^{V_m})$

   i.e. $G(.)$ is homogeneous of degree $1/\theta$, positive if $\theta > 0$;

3) $lim_{e^{V_i} \to \infty} G(e^{V_1},...,e^{V_m}) = lim_{e^{V_i} \to \infty} \sum_{i=1}^{m} e^{V_i/\theta} = \infty$, for $i = 1,2,...,m$;

4) the first derivative of $G(\cdot)$ with respect to any e^{V_j} is equal to:

$$G_k = \partial G(.)/\partial e^{V_j} = \frac{e^{V_j[(1/\theta)-1]}}{\theta}$$

it is non-negative for any $\theta \geq 0$. Furthermore, higher-order mixed derivatives are all null, and therefore both non-negative and non-positive. Condition 4) is therefore certainly verified if condition 2) on the positivity of the coefficient θ is verified.

Substituting the expression (3.A.3) in equation (3.A.1), it results:

$$p[j] = \frac{e^{V_j}}{1/\theta} \cdot \frac{1/\theta \cdot e^{V_j(1/\theta)-1]}}{\sum_{i=1}^{m} e^{V_i/\theta}} = \frac{e^{V_j/\theta}}{\sum_{i=1}^{m} e^{V_i/\theta}}$$

which is the expression of the Multinomial Logit model of parameter θ.

To complete the demonstration, the joint probability distribution of random residuals can be derived. In fact, substituting expression (3.A.3) in the joint probability distribution function (3.3.53), the product of m Gumbel probability distribution functions of parameter θ is obtained:

$$F(\varepsilon_1,...,\varepsilon_m) = exp\left[-\sum_{i=1}^{m} e^{-\varepsilon_i/\theta}\right] = \prod_{i=1}^{m} exp[-e^{-\varepsilon_i/\theta}]$$

Thus expression (3.A.3) of the function $G(.)$ implies that random residuals are identically and independently distributed as Gumbel variables of parameter θ and therefore with variances and covariances defined by expressions (3.3.2) and (3.3.3). Note that the Euler's constant Φ has been included in systematic utilities V_i with no loss of generality since, as stated in section 3.3.1, Logit choice probabilities are invariant with respect to the addition of a constant to all utilities.

3.A.2. Derivation of the Single-Level Hierarchical Logit model

In the *Single-Level Hierarchical Logit* model with equal covariances, the choice tree has the root node o connected to intermediate nodes k to which elementary alternatives j are connected (see Fig. 3.3.4). The parameters θ associated with all intermediate nodes k are equal.

With this tree structure, the variable Y_o becomes:

$$Y_o = ln\sum_k exp\left(\frac{\theta}{\theta_o} \cdot Y_k\right) = ln\sum_k \left(\sum_{i \in J_k} e^{V_i/\theta}\right)^{\theta/\theta_o}$$

with

$$Y_k = ln \sum_{i \in I_k} e^{V_i/\theta}$$

Consequently equation (3.A.2) becomes:

$$G(e^{V_1},...,e^{V_m}) = \sum_k \left(\sum_{i \in I_k} e^{V_i/\theta} \right)^{\theta/\theta_o} \tag{3.A.4}$$

Also in this case, it can be shown that $G(.)$ satisfies the four properties mentioned above, given some limitations on parameters θ and θ_o.

In fact:

1) $G \geq 0$ for any value of θ, θ_o and V_k and ($k = 1,...,m$);

2) $G(\alpha e^{V_1},...,\alpha e^{V_m}) = \sum_k \left[\sum_{i \in I_k} (\alpha e^{V_i})^{1/\theta} \right]^{\theta/\theta_o} = \sum_k \left[(\alpha)^{1/\theta} \sum_{i \in I_k} (e^{V_i})^{1/\theta} \right]^{\theta/\theta_o} =$

$= \sum_k (\alpha)^{1/\theta_o} \cdot \left[\sum_{i \in I_k} (e^{V_i})^{1/\theta} \right]^{\theta/\theta_o} = (\alpha)^{1/\theta_o} \cdot \sum_k \left[\sum_{i \in I_k} (e^{V_i})^{1/\theta} \right]^{\theta/\theta_o} =$

$= (\alpha)^{1/\theta_o} \cdot G(e^{V_1},...,e^{V_m})$

i.e. G is homogeneous of degree $1/\theta_o$, positive if $\theta_o > 0$;

3) $\lim_{e^{V_k} \to \infty} G(e^{V_1},...,e^{V_m}) = \infty$, for $k = 1,2,...,m$;

4) the first-order partial derivative of $G(.)$ with respect to any e^{V_h} is equal to:

$$G_h = \partial G(.)/\partial e^{V_h} = \theta/\theta_o \cdot \left(\sum_{i \in I_k} e^{V_i/\theta} \right)^{(\theta/\theta_o)-1} \cdot 1/\theta \cdot e^{V_h[(1/\theta)-1]} \quad \text{with } h \in I_k$$

which is non-negative if:

$$\theta_o \geq 0 \tag{3.A.5}$$

Inequality (3.A.5) is implied by condition 2) on the positivity of the homogeneity coefficient.

Moreover, second-order mixed derivatives are equal to:

$$\partial^2 G(.)/\partial e^{V_j} \partial e^{V_h} = \begin{cases} \dfrac{1}{\theta_o} \cdot e^{V_j[(1/\theta)-1]} \cdot \left(\dfrac{\theta}{\theta_o} - 1 \right) \cdot \left(\sum_{i \in I_k} e^{V_i/\theta} \right)^{(\theta/\theta_o)-2} \dfrac{1}{\theta} e^{V_h[(1/\theta)-1]} & \text{for } j,h \in I_k \ \forall k \\ 0, & \text{otherwise} \end{cases}$$

which, given (3.A.5), are non-positive if:

$$0 \leq \theta \leq \theta_o \tag{3.A.6}$$

It can be easily shown that condition 4) is always satisfied, for higher order mixed derivatives, if (3.A.6) holds.

Also in this case, therefore, conditions 2) and 4) impose limitations for the two parameters $\theta$and θ_o ($0 < \theta \leq \theta_o$) analogous to those described in section 3.3.2.

Choice probabilities can be obtained by substituting function (3.A.4) in equation (3.A.1):

$$p[j] = \frac{e^{V_j}}{\frac{1}{\theta_o}} \cdot \frac{\frac{\theta}{\theta_o} \cdot \left(\sum_{i \in I_h} e^{V_i/\theta}\right)^{\frac{\theta}{\theta_o}-1} \cdot \frac{1}{\theta} \cdot e^{V_j[(1/\theta)-1]}}{\sum_k \left(\sum_{i \in I_k} e^{V_i/\theta}\right)^{\frac{\theta}{\theta_o}}} = \frac{e^{V_j/\theta}}{\sum_{i \in I_k} e^{V_i/\theta}} \cdot \frac{\left(\sum_{i \in I_h} e^{V_i/\theta}\right)^{\frac{\theta}{\theta_o}}}{\sum_k \left(\sum_{i \in I_k} e^{V_i/\theta}\right)^{\frac{\theta}{\theta_o}}} \quad (3.A.7)$$

which is the expression of the Single-Level Hierarchical Logit model with parameters θ_o and θ. Introducing the parameter $\delta = \theta/\theta_o$ and the logsum variable Y_k:

$$Y_k = ln \sum_{i \in I_k} exp(V_i/\theta)$$

(3.A.7) becomes:

$$p[j] = \frac{e^{V_j/\theta}}{\sum_{i \in I_k} e^{V_i/\theta}} \cdot \frac{e^{\delta Y_h}}{\sum_k e^{\delta Y_k}}$$

which is another expression of the Single-Level Hierarchical Logit with constant covariances.

3.A.3. Derivation of the Multi-Level Hierarchical Logit model

The demonstration that the *Multi-Level Hierarchical Logit* (*Tree-Logit*) can be derived from function (3.A.2) satisfying the four properties mentioned cannot be easily generalized since it is difficult to express the choice tree structure in a general form. To demonstrate the statement that the Multi-Level Hierarchical Logit model is a GEV model, reference to an easily generalizable example will be made.

Consider the structure of the choice tree in Fig. 3.A.1.

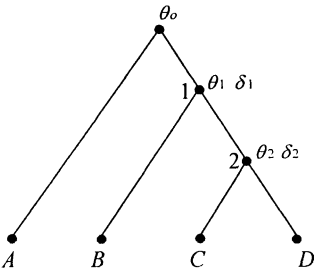


Fig. 3.A.1 Choice tree for a Multi-Level Hierarchical Logit model.

There are two intermediate levels and three parameters: θ_o, θ_1, θ_2. Let V_A, V_B, V_C and V_D be the systematic utilities of the four elementary nodes. According to what stated in section (3.3.3), it follows:

$$\begin{aligned}
&\delta_1 = \theta_1 / \theta_o \\
&\delta_2 = \theta_2 / \theta_1 \\
&Y_2 = ln(e^{V_C/\theta_2} + e^{V_D/\theta_2}) \\
&Y_1 = ln(e^{V_B/\theta_1} + e^{\delta_2 Y_2}) = ln[e^{V_B/\theta_1} + (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\theta_2/\theta_1}] \\
&Y_o = ln(e^{V_A/\theta_o} + e^{\delta_1 Y_1}) = ln\{e^{V_A/\theta_o} + [e^{V_B/\theta_1} + (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\theta_2/\theta_1}]^{\theta_1/\theta_o}\}
\end{aligned} \tag{3.A.8}$$

$$\begin{aligned}
p[A] &= \frac{e^{V_A}}{e^{Y_o}} \\
p[B] &= \frac{e^{V_B/\theta_1}}{e^{Y_o}} \cdot e^{(\delta_1 - 1)Y_1} \\
p[C] &= \frac{e^{V_C/\theta_2}}{e^{Y_o}} \cdot e^{(\delta_1 - 1)Y_1} \cdot e^{(\delta_2 - 1)Y_2}
\end{aligned} \tag{3.A.9}$$

Substituting the expression of Y_o given by (3.A.8) in (3.A.2) it results:

$$G(e^{V_A},...,e^{V_D}) = e^{V_A/\theta_o} + [e^{V_B/\theta_1} + (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\theta_2/\theta_1}]^{\theta_1/\theta_o} \tag{3.A.10}$$

It can be verified that this function satisfies the four properties given some limitations on the parameters θ.

In fact:

1) $G \geq 0$ for any value of θ_j, (j = o, 1, 2), V_i, (i = A, B, C, D);

2) $$\begin{aligned}
&G(\alpha e^{V_A},...,\alpha e^{V_D}) = (\alpha e^{V_A})^{1/\theta_o} + \{(\alpha e^{V_B})^{1/\theta_1} + [(\alpha e^{V_C})^{1/\theta_2} + (\alpha e^{V_D})^{1/\theta_2}]^{\theta_2/\theta_1}\}^{\theta_1/\theta_o} \\
&= (\alpha)^{1/\theta_o} \cdot (e^{V_A})^{1/\theta_o} + \{(\alpha)^{1/\theta_1} \cdot (e^{V_B})^{1/\theta_1} + [(\alpha)^{1/\theta_2} \cdot (e^{V_C})^{1/\theta_2} + (\alpha)^{1/\theta_2} \cdot (e^{V_D})^{1/\theta_2}]^{\theta_2/\theta_1}\}^{\theta_1/\theta_o} \\
&= (\alpha)^{1/\theta_o} \cdot (e^{V_A})^{1/\theta_o} + \{(\alpha)^{1/\theta_1} \cdot (e^{V_B})^{1/\theta_1} + (\alpha)^{1/\theta_1} \cdot [(e^{V_C})^{1/\theta_2} + (e^{V_D})^{1/\theta_2}]^{\theta_2/\theta_1}\}^{\theta_1/\theta_o} \\
&= (\alpha)^{1/\theta_o} \cdot (e^{V_A})^{1/\theta_o} + (\alpha)^{1/\theta_o} \cdot \{(e^{V_B})^{1/\theta_1} + [(e^{V_C})^{1/\theta_2} + (e^{V_D})^{1/\theta_2}]^{\theta_2/\theta_1}\}^{\theta_1/\theta_o} \\
&= (\alpha)^{1/\theta_o} \cdot G(e^{V_A},...,e^{V_D})
\end{aligned}$$

   i.e $G(.)$ is homogeneous of degree $1/\theta_o$, positive if $\theta_o > 0$;

3) $\lim_{e^{V_i} \to \infty} G(e^{V_A},...,e^{V_D}) = \infty$, for $i = A, B, C, D$;

4) first-order partial derivatives can be expressed as:

$$\partial G / \partial e^{V_A} = 1/\theta_o \cdot e^{V_A(1/\theta_o - 1)}$$

$$\partial G / \partial e^{V_B} = \theta_1 / \theta_o \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 1} \cdot 1/\theta_1 \cdot e^{V_B(1/\theta_1 - 1)}$$

$$\partial G / \partial e^{V_C} = \theta_1 / \theta_o \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 1} \cdot \theta_2 / \theta_1 \cdot (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\delta_2 - 1} \cdot 1/\theta_2 \cdot e^{V_C(1/\theta_2 - 1)}$$

Note that in this case there is no structural symmetry, and they differ from each other. First order derivatives are non-negative if:

$$\theta_o \geq 0 \tag{3.A.11}$$

Other limitations on parameters θ can be deduced from the second order mixed derivatives. In particular, it is sufficient to use only the following two mixed derivatives:

$$\partial^2 G / \partial e^{V_B} \partial e^{V_C} = \frac{1}{\theta_o} \cdot e^{V_B\left(\frac{1}{\theta_1} - 1\right)} \cdot \frac{\theta_1 - \theta_o}{\theta_o} \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 2} \cdot \frac{1}{\theta_1} \cdot (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\delta_2 - 1} \cdot e^{V_C\left(\frac{1}{\theta_2} - 1\right)}$$

$$\partial^2 G / \partial e^{V_C} \partial e^{V_D} = \frac{1}{\theta_o} \cdot e^{V_C\left(\frac{1}{\theta_2} - 1\right)} \cdot \frac{\theta_1 - \theta_o}{\theta_o} \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 2} \cdot \frac{\theta_2}{\theta_1} \cdot [(e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\delta_2 - 1}]^2 \cdot \frac{1}{\theta_2} \cdot e^{V_D\left(\frac{1}{\theta_2} - 1\right)} +$$

$$+ \frac{1}{\theta_o} \cdot e^{V_C\left(\frac{1}{\theta_2} - 1\right)} \cdot \frac{\theta_2 - \theta_1}{\theta_1} \cdot (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\delta_2 - 2} \cdot \frac{1}{\theta_2} \cdot e^{V_D\left(\frac{1}{\theta_2} - 1\right)} \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 1} \tag{3.A.12}$$

In particular the first one, imposing inequality (3.A.11), is non positive if:

$$0 \leq \theta_1 \leq \theta_o \tag{3.A.13}$$

In the second one, imposing (3.A.13), it results that the first term is always non-positive while the second term is non-positive if:

$$0 \leq \theta_2 \leq \theta_1 \tag{3.A.14}$$

Combining expression (3.A.13) and (3.A.14), it follows:

$$0 \leq \theta_2 \leq \theta_1 \leq \theta_o \tag{3.A.15}$$

It can be shown that condition 4) for the other second order mixed derivatives not included in (3.A.12) and for higher-order mixed derivatives is always verified if inequality (3.A.15) holds.

Choice probabilities for the Multi-Level Hierarchical-Logit model described, can be obtained substituting expression (3.A.10) in equation (3.A.1). It then results:

$$p[A] = \frac{e^{V_A}}{1/\theta_o} \cdot \frac{1/\theta_o \cdot e^{V_A(1/\theta_o - 1)}}{e^{Y_o}} = \frac{e^{V_A/\theta_o}}{e^{Y_o}}$$

$$p[B] = \frac{e^{V_B}}{1/\theta_o} \cdot \frac{\theta_1/\theta_o \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 1} \cdot 1/\theta_1 \cdot e^{V_B(1/\theta_1 - 1)}}{e^{Y_o}} = \frac{e^{V_B/\theta_1}}{e^{Y_o}} \cdot e^{(\delta_1 - 1)Y_1}$$

$$p[C] = \frac{e^{V_C}}{1/\theta_o} \cdot \frac{\theta_1/\theta_o \cdot (e^{V_B/\theta_1} + e^{\delta_2 Y_2})^{\delta_1 - 1} \cdot \theta_2/\theta_1 \cdot (e^{V_C/\theta_2} + e^{V_D/\theta_2})^{\delta_2 - 1} \cdot 1/\theta_2 \cdot e^{V_C(1/\theta_2 - 1)}}{e^{Y_o}} =$$

$$= \frac{e^{V_C/\theta_2}}{e^{Y_o}} \cdot e^{(\delta_1 - 1)Y_1} \cdot e^{(\delta_2 - 1)Y_2}$$

equal to the expressions (3.A.9)

The conditions on parameters θ obtained for the three models described so far are both necessary and sufficient; if they are not satisfied the function $G(.)$ doesn't have the properties 1, 2, 3 and 4 and the models are not compatible with random utility theory.

3.A.4. Derivation of the Cross-Nested Logit model

The Cross-Nested Logit model has a choice graph shown in Fig. 3.3.10 and can be obtained as a GEV model by specifying the function $G(.)$ as:

$$G(.) = \sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k} \tag{3.A.16}$$

with $\delta_k = \theta_k/\theta_o$ and the membership parameters α_{ik} in the interval [0,1]. Also in this case, it can be verified that $G(.)$ satisfies the four properties, given some limitations on parameters θ_k.

In fact:

1) $G \geq 0$ for any value of θ_k, V_i ($i = 1,...m$), a_{im} [0;1];

2) $$G(\beta e^{V_1},...,\beta e^{Vm}) = \sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} (\beta e^{V_i})^{1/\theta_k} \right)^{\delta_k} = \sum_k \left(\beta^{1/\theta_k} \sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} (e^{V_i})^{1/\theta_k} \right)^{\delta_k} =$$

$$= \beta^{1/\theta_o} \sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} (e^{V_i})^{1/\theta_k} \right)^{\delta_k} = \beta^{1/\theta_o} \cdot G(e^{V_1},...,e^{V_m})$$

i.e. $G(.)$ is homogeneous of degree $1/\theta_o$, positive if it is $\theta_o \geq 0$;

3) $\lim_{e^{V_k} \to \infty} G(e^{V_1},...,e^{V_m}) = \infty$, for $k = 1,2,...,m$;

4) the first order partial derivative of $G(.)$ with respect to any e^{V_j} is equal to:

$$G_j = \partial G(.)/\partial e^{V_j} = \sum_k \left[\delta_k \cdot \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k - 1} \cdot \frac{\alpha_{jk}^{1/\delta_k}}{\theta_k} \cdot (e^{V_j})^{\frac{1}{\theta_k} - 1} \right]$$

and is non negative if it is

$$\theta_o \geq 0; \tag{3.A.17}$$

Inequality (3.A.17) is implied by condition 2) on the positivity of the homogeneity coefficient.

Moreover, second-order mixed derivatives are equal to:

$$\partial^2 G(.)/\partial e^{V_j} \partial e^{V_h} = \sum_k \left[\alpha_{hk}^{1/\delta_k} \cdot \frac{1}{\theta_k} (e^{V_h})^{\frac{1}{\theta_k} - 1} \cdot (\delta_k - 1) \cdot \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k - 2} \cdot \frac{\alpha_{jk}^{1/\delta_k}}{\theta_o} \cdot (e^{V_j})^{\frac{1}{\theta_k} - 1} \right]$$

If inequality (3.A.17) is satisfied, all terms of the summation are non-positive if:

$$0 \le \theta_k \le \theta_o \quad \forall k \tag{3.A.18}$$

Thus the condition of non positivity is always satisfied (for any value of V_i, a_{ik}) if the (3.A.18) is true.

It can be easily shown that condition 4) for higher order mixed derivatives is always verified if (3.A.18) holds.

Choice probabilities can be obtained by substituting the function $G(.)$ expressed by (3.A.16) in equation (3.A.1):

$$p[j] = \frac{e^{V_j}}{1/\theta_o} \cdot \frac{\sum_k \left[\frac{\alpha_{jk}^{1/\delta_k}}{\theta_o} \cdot \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k - 1} \cdot (e^{V_j})^{\frac{1}{\theta_k} - 1} \right]}{\sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k}} = \tag{3.A.19}$$

$$= \frac{\sum_k \left[\alpha_{jk}^{1/\delta_k} e^{V_j/\theta_k} \cdot \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k - 1} \right]}{\sum_k \left(\sum_{i \in I_k} \alpha_{ik}^{1/\delta_k} e^{V_i/\theta_k} \right)^{\delta_k}}$$

which is the expression of the Cross-Nested Logit model (cfr. 3.3.49).

3.B. Random variables relevant for random utility models

3.B.1. The Gumbel random variable

The Gumbel variable is a continuous variable, which plays a very important role in the building of random utility models of the Logit family. In the following the probability functions of this variable are described and some important properties are illustrated. To facilitate the immediate application of the results to random utility models, the Gumbel variable will be indicated by U (instead of X_G) and its expected value by V (instead of $E[X_G]$).

The probability density function of a Gumbel r.v. U with mean V and parameter θ is given by:

$$f_U(u) = 1/\theta \cdot exp[-(u-V)/\theta - \Phi] \; exp\{-exp[-(u-V)/\theta - \Phi]\} \qquad (3.B.1)$$

and its distribution function is:

$$F_U(u) = exp\{-exp[-(u-V)/\theta - \Phi]\} \qquad (3.B.2)$$

where Φ is the Eulero's constant approximately equal to 0.577.

The mean and the variance of the Gumbel variable are:

$$E[U] = V$$
$$Var[U] = \sigma^2_U = \frac{\pi^2\theta^2}{6} \qquad (3.B.3)$$

From expressions (3.B.3) it is deduced that the standard deviation of the Gumbel r.v. is directly proportional to the parameter θ. Fig. 3.B.1 shows some probability density functions of the Gumbel r.v. with zero mean for different parameters θ.

It can easily be demonstrated, by substitution in expression (3.B.2), that if U is a Gumbel variable with parameters (V,θ), any r.v. obtained by linear transformation

$$Y = aU + b$$

is also a Gumbel r.v. of mean

$$E[Y] = aV + b$$

and the same parameter θ (same variance). From this result it follows immediately that the residual of a random utility model $\varepsilon = U-V$ ($a = 1$, $b = -V$) is a Gumbel r.v. of zero mean and parameter θ.

The Gumbel r.v. has an important property of *stability with respect to maximization*. In other words, if U_j, $j=1,...,N$, are independent Gumbel r.v. with different mean V_j but the same parameter θ, the maximum of these variables:

$$U_M = max_{j=1,\ldots,N}[U_j] \tag{3.B.4}$$

is also a Gumbel r.v. of parameter θ.

In fact, the probability distribution function of U_M, , can be obtained as:

$$F_{U_M}(u) = Pr(U_M < u) = Pr[max_{j=1,\ldots,N}\{U_j\} \leq u]$$

and for the independence of the U_j, it follows:

$$Pr[max_{j=1,\ldots,N}\{U_j\} \leq u] = \prod_{j=1,\ldots,N} Pr[U_j < u] = \prod_{j=1,\ldots,N} F_{U_j}(u)$$

Substituting the expression (3.B.2) of the Gumbel probability distribution function in the previous expression, it follows that:

$$F_{U_M}(u) = \prod_{j=1,\ldots,N} exp\{-exp[-(u - V_j)/\theta - \Phi]\}$$

which yields:

$$F_{U_M}(u) = exp[-exp(-\Phi) \cdot exp(-u/\theta) \cdot \sum_j exp(V_j/\theta)] \tag{3.B.5}$$

If the EMPU variable described in Chapter 3 is indicated by V_M:

$$V_M = \theta \, ln \, \Sigma_j \, exp(V_j/\theta) \tag{3.B.6}$$

and it is substituted in the expression (3.B.5), it follows

$$F_{U_M}(u) = exp\{-exp[-(u - V_M)/\theta - \Phi]\}$$

which is still the probability distribution function of a Gumbel random variable with mean V_M and parameter θ as is deduced immediately by comparison with (3.B.2).

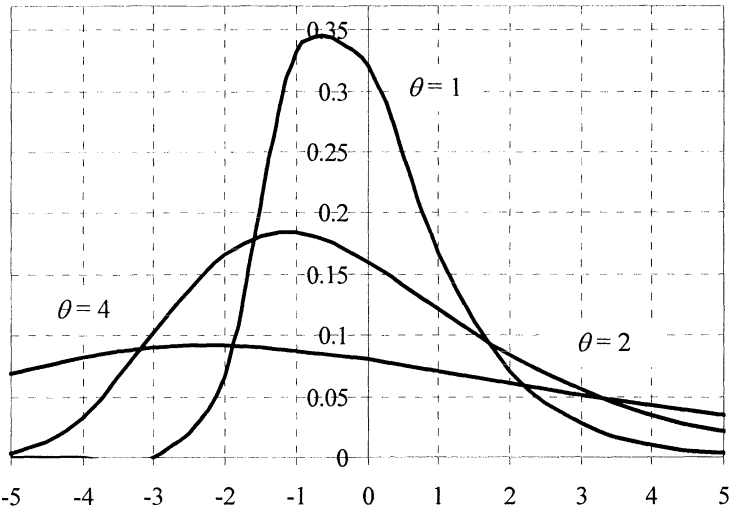


Fig. 3.B.1 Probability density functions of a Gumbel r.v.

The Multinomial Logit model can be obtained by using the definition of random utility model (3.2.1) and the property of stability with respect to maximization of the Gumbel r.v. described above.

In fact, from (3.2.1) it results:

$$p[j] = Pr(U_j > U_{M'})$$

with

$$U_{M'} = max_{k \neq j}\{U_k\}$$

This probability can therefore be expressed as the product between the probability that the perceived utility U_j has a value included in an infinitesimal neighborhood of x and the probability that U_M has a value less than x. This product must obviously be integrated with respect to all the possible values of x:

$$p[j] = Pr(U_j > U_{M'}) = \int_{-\infty}^{+\infty} F_{U_{M'}}(x) \cdot f_{U_j}(x)\,dx \qquad (3.B.7)$$

where $F_{U_{M'}}$ and f_{U_j} are the probability distribution function and the probability density function of the random variables $U_{M'}$ and U_j respectively. If the U_k are i.i.d. Gumbel variables of parameter θ and mean V_k, $U_{M'}$, as shown above, is also a Gumbel variable with the same parameter θ and mean equal to:

$$V_{M'} = \theta\, ln\, \Sigma_{k \neq j}\, exp\,(V_k/\theta) \qquad (3.B.8)$$

Expression (3.B.7) then becomes:

$$p[j] = \int_{-\infty}^{+\infty} exp\{-exp[-(x - V_{M'})/\theta - \Phi]\} \cdot exp\{-exp[-(x - V_j)/\theta - \Phi]\} \cdot$$
$$\cdot exp[-(x - V_j)/\theta - \Phi] \cdot (1/\theta)\, dx =$$

$$= \int_{-\infty}^{+\infty} exp\{-exp[-(x - V_j)/\theta - \Phi] - exp[-(x - V_{M'})/\theta - \Phi]\} \cdot$$
$$exp[-(x - V_j)/\theta - \Phi] \cdot (1/\theta)\, dx =$$

$$= exp(V_j/\theta - \Phi) \cdot \int_{-\infty}^{+\infty} exp\{-exp(-x/\theta) \cdot [exp(V_j/\theta - \Phi) + exp[V_{M'}/\theta - \Phi]\} \cdot$$
$$exp(-x/\theta) \cdot (1/\theta)\, dx =$$

$$= exp(V_j/\theta - \Phi) \cdot \int_{-\infty}^{+\infty} exp[-exp(-x/\theta)]^{[exp(V_j/\theta - \Phi) + exp[V_{M'}/\theta - \Phi]}$$
$$\cdot exp(-x/\theta) \cdot (1/\theta)\, dx =$$

$$= \frac{exp(V_j/\theta - \Phi)}{exp(V_j/\theta - \Phi) + exp(V_{M'}/\theta - \Phi)} \cdot \left| exp[-exp(-x/\theta)]^{[exp(V_j/\theta - \Phi) + exp[V_{M'}/\theta - \Phi]} \right|_{-\infty}^{+\infty} =$$

$$= \frac{exp(V_j/\theta)}{exp(V_j/\theta) + exp(V_{M'}/\theta)}$$

and, substituting expression (3.B.8) for $V_{M'}$, it follows:

$$p[j] = \frac{exp(V_j/\theta)}{exp(V_j/\theta) + \sum_{k \neq j} exp(V_k/\theta)} = \frac{exp(V_j/\theta)}{\sum_k exp(V_k/\theta)}$$

which is the Multinomial Logit model described in section 3.3.1.

3.B.2. The Multivariate Normal random variable

The Multivariate Normal r.v., X_{MVN}, is the generalization of the normal r.v. to the case of n dimensions. Its probability density function is given by:

$$f_{X_{MVN}}(\boldsymbol{x}) = [2\pi^n det(\Sigma_X)]^{-1/2} exp[-1/2(\boldsymbol{x} - \mu_X)^T \Sigma_X^{-1}(\boldsymbol{x} - \mu_X)] \qquad (3.B.9)$$

where the parameters are the vector μ and the matrix Σ, and $det(\Sigma)$ denotes the determinant of the matrix Σ.

The parameters of a multivariate normal r.v. are the vector of the means with components μ_{X_i} and the variance-covariance matrix that can therefore assume any value as long as it is positive semi-definite. In other words:

$$E[X_{MVN}] = \mu_X \qquad \Sigma_{X_{MVN}} = \Sigma_X$$

The surfaces of equiprobability of the multivariate normal variable, or the loci of points in the n-dimensional Euclidean space for which the density function is constant, have the equation:

$$(\boldsymbol{x} - \mu_X)^T \Sigma_X^{-1} (\boldsymbol{x} - \mu_X) = C^2 \qquad (3.B.10)$$

where C is a constant. Expression (3.B.10) is the equation of an ellipsoid with μ as its center (see Fig. 3.B.2).

The Multivariate normal r.v. has the property of invariance with respect to linear transformations, which can be considered an extension of the property of invariance with respect to the summation of the normal r.v. In other words, if $\boldsymbol{X}$ is a random vector with probability density function (3.B.9), the vector $\boldsymbol{Y} = \boldsymbol{AX}$, where $\boldsymbol{A}$ is a matrix of dimensions $(m \times n)$, is also a Multivariate normal variable with mean vector and dispersion matrices given by:

$$E[\boldsymbol{Y}] = \boldsymbol{A}E[\boldsymbol{X}] = \boldsymbol{A}\,\mu_X$$
$$\Sigma_Y = E[\boldsymbol{A}(\boldsymbol{X} - \mu_X)(\boldsymbol{X} - \mu_X)^T \boldsymbol{A}^T] = \boldsymbol{A}\,\Sigma_X \boldsymbol{A}^T$$

Furthermore, from (3.B.9) it can be easily deduced that if the n components of X_{MVN} are non-correlated, i.e. the matrix Σ is diagonal, they are independent, i.e. the probability density function (3.B.9) is the product of n density functions of

univariate normal r.v. with means μ_{X_i} and variances $\sigma^2_{X_i}$. It can be worth recalling that two independent random variables are non-correlated in any case.

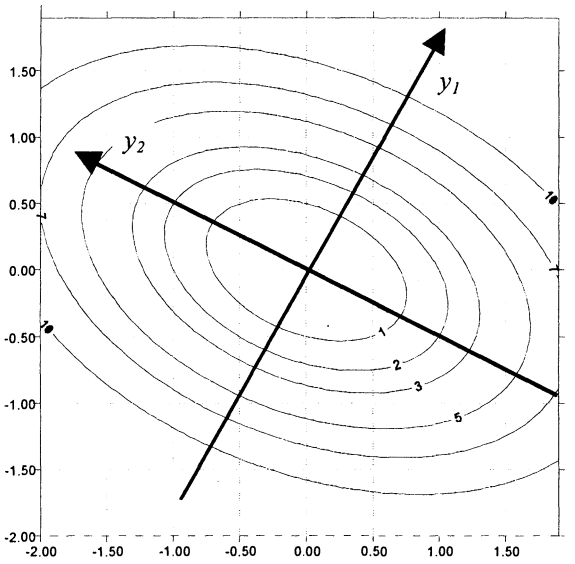


Fig. 3.B.2 Equiprobable surfaces of the Multivariate Normal r.v.

Reference Notes

Random utility theory has stimulated, both in theory and in applications, the understanding and modeling of the mechanisms underlying transportation demand. One of the first systematic accounts of its foundation can be found in the book by Domencich and McFadden (1975). The book formalizes the theoretical work carried out in the early 70's on random utility models and on Multinomial Logit models in particular.

Theoretical analyses of random utility models can be found in Williams (1977), Manski (1977) and in the book by Manski and McFadden (1981). The book by Ben-Akiva and Lerman (1985), gives a very comprehensive account of random utility theory, of Logit family models and of many applicative issues dealt with in this chapter (e.g. elasticities, aggregation procedures) and in Chapter 8 (calibration and tests).

The work of Williams and Ortuzar (1982) analyzes the limitations of random utility (or "compensatory") models and compares them with other behavioral discrete choice models. The paper also contains a comprehensive, though dated, bibliography on non-compensatory models. A detailed analysis of the state of the art on the use of random utility models in modeling transportation demand can be found

in the note by Horowitz (1985). More recent systematic reviews of random utility models can be found in Bath (1997) and in Ben-Akiva and Bierlaire (1999).

As for specific random utility models, references to the Single-Level Hierarchical Logit model can be found in Williams (1977) and Daly and Zachary (1978), while for an exposition of the Multi-Level Hierarchical Logit model in its most general form, reference can be made to Daganzo and Kusnic (1992). The Multi-Level Hierarchical-Logit have been reformulated and specified in more compact and general form than in the literature. The Cross-Nested Logit model was first presented by McFadden (1978) in the context of residential location choice; applications to other choice dimensions ca be found in Small (1987) and Vovsha (1997) (1998). The formulation reported in section 3.3.4 is from Papola (2000). The GEV model was proposed by McFadden (1978) and subsequently generalized by Ben-Akiva and Francois (1983). The demonstration of GEV models as random utility models and the derivation of Hierarchical Logit models as GEV models is from Papola (1996) while the derivation of the Cross-Nested Logit model as a GEV model is from Papola (2000).

A detailed analysis of the Probit model is contained in the book by Daganzo (1979); for the calculation of Probit choice probabilities reference can be made to the work of Horowitz, Spermann and Daganzo (1982) and Langdon (1984). Reference to the Factor Analytic Probit can be found in Ben-Akiva and Bierlaire (1999) while reference to the Random Coefficients (Tastes) approach can be found in Ben-Akiva and Lerman (1985) and in Ortuzar and Willumsen (1994).

The Hybrid Logit-Probit model is also a rather recent development of random utility models. One of the first papers dealing with its theoretical and computational aspects is Ben-Akiva and Bolduc (1996). Other references to this model are in Bolduc et al. (1996) and in Ben-Akiva and Bierlaire (1999).

The general approach to modeling choice set alternatives is contained in Manski (1977). A recent outline of the state of the art of explicit models of choice set generation and a number of specifications are found in Ben-Akiva and Boccara (1995). The Implicit Availability Perception approach is described in Cascetta and Papola (2000).

The Expected Maximum Perceived Utility function and its mathematical properties are dealt with in Daganzo's volume (1979). Reference can also be made to the work of Cantarella (1997) which draws on and generalizes the results.

The definition of elasticity associated with random utility models and the expressions for the Multinomial Logit model are given in various texts, particular reference can be made to Domencich and McFadden (1975) and to Ben-Akiva and Lerman (1985). The results relative to elasticities for the Single-Level Hierarchical Logit model are from Koppelman (1989).

A systematic treatment of the aggregation procedures proposed by Ben-Akiva and Atherton (1977) is given in Ben-Akiva and Lerman (1985). The "prototypical" sample method is described in Watanatada and Ben-Akiva (1979) and in Gunn and Bates (1982).

Notes

[1] Behavioral models, like all microeconomic demand models, attempt to reproduce the results of choice behavior "as if" decision makers behaved in accordance with certain hypotheses rather than the actual psychological mechanism leading to decisions.

[2] Discrete choice models in general, and random utility models in particular, can be considered one of the most significant contributions of the transport field to economis and econometrics. From the theoretical point of view, they represent the development of the classical micro-economic demand models. In fact discrete choice models simulate choices made among discrete alternatives while classical micro-economic demand models simulate the choice of the (continuous) quantity of "commodities" to be consumed. Discrete choice models, originally developed to simulate transport demand, are used in many fields of econometrics, from the choice of insurance policies type and investment portfolios to the choice of car models.

[3] The case in which the variance-covariance matrix is non-null, $\Sigma \neq 0$, but singular, $|\Sigma| = 0$, because the variance of a random residual is null and/or two random residuals are perfectly correlated, is of limited practical interest and will not be given further attention.

[4] This denomination derives from early applications of random utility models to the choice among different transport modes.

[5] In this section, for the sake of simplicity, the symbol i indicating the generic decision maker will be systematically given as understood.

[6] Some texts, assume as the Gumbel parameter the reciprocal of θ, i.e. $\alpha = 1/\theta$. In the following the θ notation will normally be used for its analytical convenience in the specification of Hierarchical Logit models. Clearly it is possible to express all results using the parameter α with a simple variable substitution.

[7] The Expected Maximum Perceived Utility variable will be dealt with extensively in section 3.5.

[8] Stability with respect to maximization of the Gumbel variable and the derivation of the Multinomial Logit model from the general expression (3.2.3) are described in Appendix 3.B.

[9] This property and its implications hold for the whole class of additive models, as was stated in section 3.2. In the following, the general results will be particularized for the Logit model, where they can be obtained analytically.

[10] The Hierarchical Logit model is also known in the international literature as Nested Logit .

[11] The Hierarchical Logit model can be obtained in a different and more rigorous way, as a special case of the GEV model described in section 3.3.5.

[12] The r.v. τ^*_k, for the stability with respect to maximization of the Gumbel variable, is distributed like each variable $\tau_{j/k}$ associated to alternatives j belonging to group k, i.e. as a Gumbel variable of zero mean and parameter θ.

[13] Cross Nested Logit models have been less studied with respect to other models of the Logit family. In this section a single-level cross-nesting structure will be presented. From this point of view, the Cross-Nested Logit model discussed is a generalization of the Single-Level Hierarchical Logit.

[14] Further elements of the Multivariate Normal random variable are given in Appendix 3.B.

(15) The Hybrid Logit-Probit model imposes an upper bound on the correlation of any pair of random residuals due to the variance σ^2 of i.i.d. Gumbel residuals. In fact the maximum correlation between two alternatives is:

$$Corr(\varepsilon_j, \varepsilon_h) = \frac{Cov[\varepsilon_j, \varepsilon_h]}{[Var[\varepsilon_j]]^{1/2} \cdot [Var[\varepsilon_h]]^{1/2}} = \frac{[Var[\xi_j]]^{1/2} \cdot [Var[\xi_h]]^{1/2}}{[Var[\xi_j] + \sigma^2]^{1/2} \cdot [Var[\xi_h] + \sigma^2]^{1/2}}$$

(16) The Binomial Logit model (3.4.6) should be seen as a functional relationship rather than a random utility model since it does not simulate any "choice".

(17) This consideration clarifies the importance of information on the availability of alternatives.

(18) In what follows, for the sake of simplicity, the dependence of the Expected Maximum Perceived Utility on the joint density function $f(\varepsilon)$ and on choice set I^i is not expressed. When the choice set is not observed, Expected Maximum Perceived Utility should be calculated by averaging over the various choice sets with their probabilities. The index i denoting the generic decision-maker will also be given as understood.

(19) The maximum of i.i.d. Gumbel variables with zero mean and variance σ^2 is also a Gumbel variable with the same variance. See also Appendix 3.B.

(20) The availability of a new alternative can be seen, in fact, as a passage of the systematic utility of that alternative from minus infinity to a finite value.

(21) Convexity of a scalar function of a vector is defined in Appendix A.

(22) Deterministic utility models and their properties will mainly be used in section 4.3.4 on path choice models and in Chapter 5 on assignment models.

(23) Property 1) requires the introduction of the concept of subgradients of a convex function.

(24) The actual number of decision-makers with the same attributes actually choosing alternative j is a random variable, so is the ratio between this number and the total number of decision-makers. The mean of this r.v. is equal to choice probability $p[j]$ given by the model.

(25) The elasticities discussed in this section are disaggregate, i.e. related to variations in the probabilities of a single decision maker or of a group of decision makers sharing the same values of the attributes. Aggregate elasticities refer to variations in the average choice fraction:

$$\bar{p}(j) = \sum_{i=1}^{n} p^i(j)$$

in a group of decision makers with different attributes. Variations are computed with respect to an uniform infinitesimal variation of a given attribute. In this case it is possible to express the aggregate elasticity as a weighed average of individual elasticities. For instance the direct point elasticity will be:

$$E_{kj}^{\bar{p}[j]} = \frac{\sum_{i=1}^{n} p^i[j] E_{kh}^{p^i[j]}}{\sum_{i=1}^{n} p^i[j]}$$

(26) The increasing monotonicity of Multinomial Logit choice probabilities with respect to systematic utilities is obtained again. It holds, more in general, for all additive models described in previous sections.

[27] Further elements of sample theory are discussed in Chapter 8 on demand estimation and its bibliography.

4 TRANSPORTATION DEMAND MODELS

4.1. Introduction

Recall from Chapter 1 that transportation demand derives from the need to carry out activities in different locations. Thus, its level and characteristics are influenced by the activity system and the transportation supply in the area.

In order to analyze and design transportation systems, it is necessary to estimate the present demand and to simulate its variations due to the projects under study and/or to variations in external factors. Mathematical demand models can be used for all these purposes.

A *transportation demand model*[1] can be defined as a mathematical relationship associating the average values of demand flows with their relevant characteristics to given activity and transportation supply systems. In formal terms, it can be expressed as follows:

$$d_{od}\,[K_1, K_2, \ldots] = d(\boldsymbol{SE}, \boldsymbol{T}; \boldsymbol{\beta}) \qquad (4.1.1)$$

where the average travel demand flow between zones o and d with characteristics K_1, K_2,..., is expressed as a function of a vector, $\boldsymbol{SE}$, of socio-economic variables related to the activity system and/or to the decision makers and of a vector $\boldsymbol{T}$ of level-of-service attributes of the transportation supply system, typically obtained by the models described in Chapter 2[2]. Demand functions also depend on a vector $\boldsymbol{\beta}$ of coefficients or parameters[3].

Each trip is the result of several choices made by the users, i.e. the individual traveler or the operators (manufacturers, shippers and carriers) in freight transport. In the case of the traveler, these choices range from long-term decisions such as residence and employment location and vehicle ownership, to more frequent decisions such as trip frequency, timing, destination, mode and route. In freight transport, long-term decision influencing transport demand include the location of production plants and acquisition/selling markets, the ownership, of a fleet of freight vehicles, storage facilities etc. Short-term decisions include shipment frequency, choice of mode, intermodal operator, route etc. The choices underlying a journey are made with respect to different *choice dimensions;* these are defined by a set of available alternatives and by the values of their relevant attributes. For example, the mode choice dimension is defined by the alternative transport modes available for a

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

given origin-destination pair together with their attributes. In the same trip, the user may also make choices on other dimensions, such as route and destination.

The literature proposes several mathematical models to simulate travel demand[4]; these models are based on different assumptions and have different specifications. Before describing some such model families in detail, some classification criteria will be introduced (see Fig. 4.1.1).

TYPE OF SIMULATED CHOICES	Mobility or context models
	Travel models
APPROACH TO TRAVEL REPRESENTATION	Trip demand models
	Journey or trip chaining models
	Activity participation models
BASIC ASSUMPTIONS	Behavioral models
	Descriptive models

Fig. 4.1.1 Classification factors of travel demand models.

The first classification factor is related to the type of choice (i.e. choice dimension) implicitly or explicitly simulated by the model. Decisions on some choice dimensions influence individual trips indirectly, by identifying the context or the conditions. Locations of residence and workplace, holding a driving license and the number of cars owned by the household are examples of this type of dimension. Residence and workplace locations determine the origin and destination of commuting trips, holding a license makes the car available as a transport mode, and so on. These choice dimensions and the models simulating them are known as *mobility choices* and *models.*

Usually, mobility choices are relatively stable over time since they have high variation costs and can be assumed invariant in the short term. *Travel choices* and *models* relate to dimensions characterizing individual journeys (sequences of trips) and/or the trips comprising them. Frequency, destination, transport mode, and route are examples of this type of choice dimension.

The second classification factor relates to the approach taken for simulating travel demand, i.e. the reciprocal conditioning of decisions (choices). *Trip demand models* implicitly assume that the choices relating to each origin-destination trip are made independently of choices for other trips within the same journey and other journeys. This approximation is used to simplify the analysis. This assumption is reasonable when most of the journeys in the reference period consist of two trips (origin – destination – origin), also called round trip journeys.

Journey demand or *trip chaining models*, on the other hand, assume that the choices concerning the entire journey influence each other. In this case, the intermediate destination, if any, is chosen taking into account preceding or subsequent destinations, transport modes taking into account the whole sequence of trips, and so on. Models of this type have been studied for several years and are

applied to real contexts, though less frequently than trip demand models. Some examples of models of this type will be described in section 4.4.

Finally, *activity participation demand models* simulate transportation demand as the result of the need to participate in different activities in different places. They therefore take into account the relationships occurring among different journeys made by the same person during an "average" day and, in the most general case, between journeys made by the members of the same household. Models of this type are obviously more complex than those described previously and are aimed at understanding relationships between the demand for travel and the organization of the different activities of a person and his/her family. These models are presently at the research stage and will not be dealt with in this chapter.

The last classification factor considered relates to the basic assumptions of the models. Models are known as *interpretative* or *behavioral* if they derive from explicit assumptions about users' choice behavior and *non-interpretative* or *descriptive* if they describe the *relationships between* travel demand and the variables ***SE*** and ***T*** without making specific assumptions about decision makers' behavior. There are also *mixed model systems* in which some sub-models are behavioral while others are descriptive[5].

Models of all types are also classified as either *aggregate* or *disaggregate.* In the first case, the variables (attributes) refer to a group of users (e.g., average times and monetary costs of all the trips between two traffic zones or the average number of cars owned by families of a certain category). In the second case, the variables refer to the individual user (e.g., times and costs between actual origin and destination points or the number of cars in the household). The level of variable aggregation depends on the purpose of demand modeling. The prevailing use considered in this book is simulation of the whole transport system, schematized through a network model. This implies an aggregation level which is at least zonal since, as stated in Chapters 1 and 2, the level-of-service variables ***T*** obtained with network models relate to pairs of centroid nodes representing traffic zones[6].

Finally, it should be noted that transportation demand models, like all models used in engineering and econometrics, are schematic and simplified representations of complex real phenomena intended to quantify certain relationships between the variables relevant to the problem under study. They should not be expected therefore to give a "perfect" reproduction of reality especially when this is largely dependent on individual behavior, as is the case with transportation demand. Furthermore, as will be seen later, different models with different levels of accuracy and complexity can describe the same context. However, more sophisticated models require more resources (data, specification and calibration, computing time, etc.) which must be justified by the requirements of the application.

The sections of this chapter will describe the characteristics of different types of transportation demand models with an emphasis on passengers travel demand. Section 4.2 will describe the partial share systems of trip demand models. Individual sub-models, including emission (or frequency), distribution, modal choice and route choice, as well as an example of an overall model system for intercity travel, are

described in section 4.3. Section 4.4 deals with some trip chaining demand models. Section 4.5 discusses the interpretation of results obtained with demand models and their application for different purposes. Finally, section 4.6 describes some models used for the simulation of freight transportation demand.

4.2. Trip demand model systems

Trip demand models simulate the average number of trips of given characteristics undertaken in a specific reference period (average trip flows). The trip characteristics often considered relevant are:

- s the purpose, or more properly the pair of purposes of the trip[7];
- h the period, i.e. the time band in which trips are undertaken;
- o,d the zone of origin and of destination of the trip;
- m the mode, or sequence of modes, used during the trip;
- k the route used for the trip, represented by a series of links connecting the centroids o and d on the network representing the transport service supply of mode m.

Furthermore, demand models are often differentiated by user groups homogeneous with respect to their attributes, parameters and the functional form of the models themselves. Such user groups are usually known as *user categories*. The generic user category is denoted by i. This formulation does not preclude the possibility of having disaggregate models and aggregation procedures based on methods of sample enumeration (see section 3.7) since the categories can be considered coincident with single individuals. In this case i denotes the generic individual.

With demand flow denoted by $d^i_{od}\,[s, h, m, k]$, the demand model is formally expressed as:

$$d^i_{od}\,[s, h, m, k] = d(\boldsymbol{SE}, \boldsymbol{T}) \tag{4.2.1}$$

Assuming that the decision maker is in zone o, the choice dimensions involved are typically travel choices: the number of trips (x) for purpose s, the time period h, the destination (d), the transport mode (m) and the route (k). Although travel choices are dependent on each other, for reasons of analytical and statistical convenience it is usually preferable to "decompose" the global demand function into the product of sub-models, each of which relates to one or more choice dimensions.
The sequence most often used is the following:

$$d^i_{od}[s,h,m,k]=n^i[o]\Sigma_x p^i[x/osh](\boldsymbol{SE},\boldsymbol{T})\cdot p^i[d/osh](\boldsymbol{SE},\boldsymbol{T})\cdot p^i[m/oshd](\boldsymbol{SE},\boldsymbol{T})\cdot p^i[k/oshdm](\boldsymbol{SE},\boldsymbol{T}) \tag{4.2.2}$$

where:

$n^i[o]$	is the number of individuals belonging to category i in zone o;
$p^i\,[x/osh]$ (***SE***, ***T***)	is the trip emission or frequency model, which gives the fraction of category i users who, being in o, undertake x trips for purpose s in the reference period h;
$p^i\,[d/osh]$ (***SE***, ***T***)	is the distribution model, which gives the fraction of category i users who, undertaking a trip from o for purpose s in the period h, travel to destination zone d;
$p^i\,[m/oshd]$ (***SE***, ***T***)	is the modal choice or split model, which gives the fraction of category i users who, traveling between o and d for purpose s in the period h, use transport mode m;
$p^i\,[k/oshdm]$ (***SE***, ***T***)	is the route choice model, which gives the fraction of category i users who, traveling between o and d for purpose s in the period h by mode m, use route k.

The system of models described above simulates the average trip demand flow with its relevant characteristics by initially estimating the total number of trips (*demand level*) from each zone o in the reference period $d_o[sh]$ and then splitting these trips between the possible destinations, modes and routes. For this reason the model is known as a *partial share model* (or system of models). The model described also assumes that destination, mode and path fraction model do not depend on the number of trips undertaken.

The order of the sequence of sub-models in expression (4.2.2) may differ from that described. Each formulation (or specification) corresponds to an assumption about the order in which the choices corresponding to different dimensions are made by the user and therefore about how they influence each other. The specification used in (4.2.2), which corresponds to the structure of models shown in Fig. 4.2.1, implies for example that mode choice depends on destination and frequency choices, while route choice depends on mode choice. On the other hand, upper-level choices (e.g. destination) are actually made taking into account the alternatives available at lower-levels, such as the modes and routes available to reach each destination. Different sequences are clearly possible therefore; for example some specifications proposed in the literature invert the position of destination and mode in the sequence (4.2.2). Any sequence should be carefully reviewed in the calibration phase (see Chapter 8) and compared with reasonable alternatives.

The fractions included in the models may be different from those shown; expression (4.2.2), because of its structure, is also known as the *four-level* or *four-stage model*. However, more or less levels can be used. For example, it is possible to specify a six-level urban demand model which explicitly includes a choice model for the time period h in which to conduct a given activity (trip purpose s) $p^i[h/osx]$ (***SE***, ***T***) and a choice model of parking location (d_p) and type (t_p) for auto trips (a) between origin o and final destination d, $p^i[d_p t_p/oshda]$(***SE***,***T***):

$$d^{i}_{od}[s,h,a,t_p,d_p,k]=n^{i}[o]\cdot\sum_{x}\left\{p^{i}[x/os](\mathbf{SE},\mathbf{T})\cdot\sum_{y_h=1}^{x}y_h\cdot p^{i}[y_h/osx](\mathbf{SE},\mathbf{T})\right\}\cdot$$

$$p^{i}[d/osh](\boldsymbol{SE},\boldsymbol{T})\cdot p^{i}[a/oshd](\boldsymbol{SE},\boldsymbol{T})\cdot p^{i}[d_pt_p/oshda](\boldsymbol{SE},\boldsymbol{T})\cdot p^{i}[k/oshdad_pt_p](\boldsymbol{SE},\boldsymbol{T}) \quad (4.2.3)$$

where x in this case represents the number of trips undertaken in a longer period (e.g. the whole day) and y_h represent the number of trips undertaken in the time period h (hourly time bands) given x.

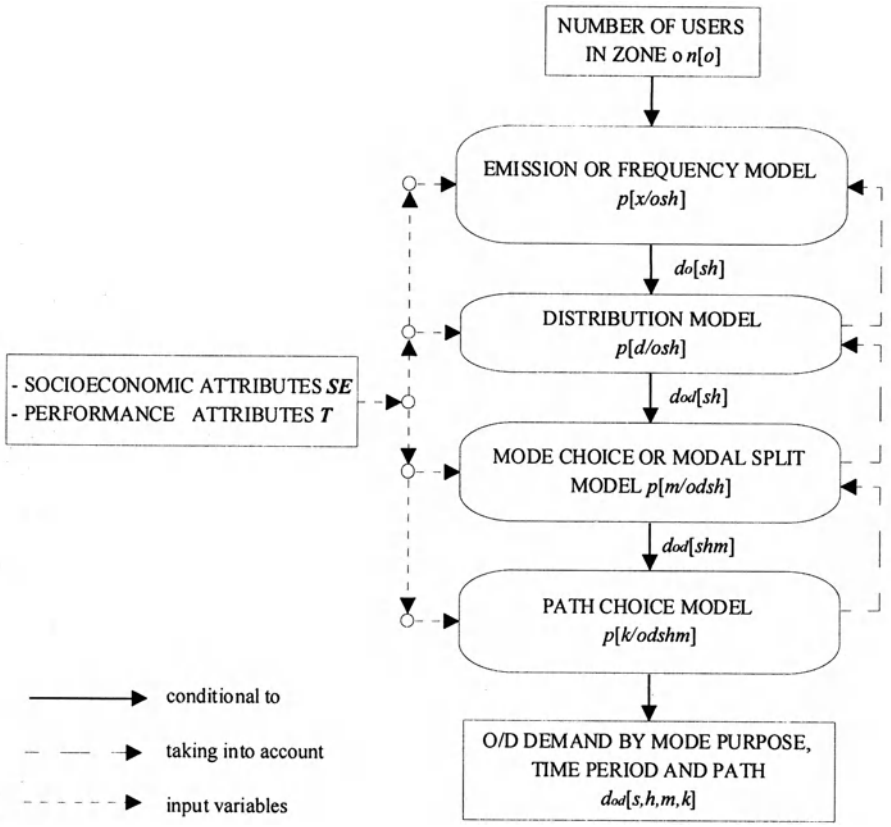


Fig. 4.2.1 Four-level trip demand model system.

The model structures described simulate trip demand over all choice dimensions. This is common practice if the projects planned and/or the evaluation time horizon significantly modifies performance and/or activity variables. In some short-term project applications, a "reduced" version of the model described could be used, e.g. assuming as given present origin-destination matrices by purpose and user category $d^{i}_{od}[sh]$ and simulating only mode and path choice levels:

$$d^{i}_{od}[s, h, m, k] = d^{i}_{od}[sh] \cdot p^{i}[m/oshd](\boldsymbol{SE}, \boldsymbol{T}) \cdot p^{i}[k/oshdm](\boldsymbol{SE}, \boldsymbol{T}) \qquad (4.2.4)$$

Estimates of present O-D matrices $d^{i}_{od}[sh]$ can be obtained with different methods as will be seen in detail in Chapter 8.

Random utility models for partial shares. Each partial share in the previous structure can be modeled in many ways. However, it is particularly interesting to express systems of partial share models consistent with the general results of random utility theory reported in Chapter 3. Again, each trip can be seen as the result of choices made on several dimensions. Therefore, choice alternatives of a random utility model are combinations of alternatives in the dimensions considered. For example, an alternative in a four-level model is defined by the number of trips x to be made for purpose s, in the reference period h, in order to reach destination d, by mode m and path k. In this case the symbol j denoting the generic alternative in Chapter 3, is equivalent to the sequence $[x,d,m,k]$. To simplify the notation, i will be understood as being the user class, o the origin zone, s the trip purpose and h the reference period.

Thus the problem is that of making consistent assumptions on the structure of the utility of compound alternatives $[x,d,m,k]$ such that the probability of choosing any of them can be expressed as the product of conditional choice probabilities representing the individual shares. This problem can be defined as *factorialization of random utility models.*

Consider each of the models expressing a different partial share as a random utility model, possibly of a different functional form (e.g., Multinomial Logit for mode choice, Probit for car route choice and deterministic for transit and walk path choice). In this case the entire demand model turns out to be consistent with the behavioral assumptions of random utility theory if the systematic utility corresponding to each choice dimension includes, as utility attribute, the EMPU corresponding to the choice made on the "lower" dimension (and, through this, those on the lower dimensions).

Consider first a very simple example for a two-level model. The model assumes there is a single mode and consists only of destination and path choice models. The perceived utility U_{dk} of the compound alternative destination-path dk is given by:

$$U_{dk} = V_{dk} + \varepsilon_{dk} \qquad (4.2.5)$$

Assume that the systematic utility can be expressed as the sum of two terms: $E[U_{dk}]=V_{dk}=V_d+V_{k/d}$, where V_d includes the attributes depending only on the destination and $V_{k/d}$ those related to the path for a given destination. Similarly the random residual is expressed as the sum of two independent random variables: $\varepsilon_{dk}=\tau_{k/d}+\eta_d$, thus yielding:

$$U_{dk} = V_d + V_{k/d} + \tau_{k/d} + \eta_d \qquad (4.2.6)$$

Clearly, the variance of any random residual ε_{dk} is the sum of the variances of $\tau_{k/d}$ and η_d.

These assumptions on the utility structure are consistent with the hierarchy of choice dimensions described before: the choice of destination is influenced by that of path but the latter, for a given destination, depends only on the attributes of the alternative paths and not on those specific to the destination.

The perceived utility for a path k, conditional to destination d, is given by:

$$U_{k/d} = V_{k/d} + \tau_{k/d}$$

since V_d and η_d are constant for all paths k, $k \in K_{od}$. The conditional probability of choosing path k for a given destination d can thus be expressed as:

$$p[k/d]=Pr[V_{k/d}+ \tau_{k/d}> V_{k'/d} + \tau_{k'/d}] \; \forall \; k' \in K_{od} \tag{4.2.7}$$

If random residuals $\tau_{k/d}$ for path k conditional to destination d are jointly distributed as a MVN variable, a Probit model is obtained for path choice.

The perceived utility for a destination d, whichever the path, is given by:

$$U^*_d = V_d + max_k\,(V_{k/d} + \tau_{k/d}) + \eta_d \tag{4.2.8}$$

The random variable $max_k\,(V_{k/d} + \tau_{k/d})$ can be expressed by the sum of its expected value $E[max_k\,(V_{k/d} + \tau_{k/d})] = s_d(V_{k1/d}, V_{k2/d},...)$ (EMPU) plus the random residual[8] τ^*_d, thus yielding:

$$U^*_d = V_d + s_d(V_{k1/d}, V_{k2/d},...) + \tau^*_d + \eta_d \tag{4.2.9}$$

It must be remembered that the EMPU variable s_d is a function whose value depends on the values of the systematic utilities of all the alternative paths, $k \in K_{od}$, as well as on the joint probability function of the random residuals $\tau_{k/d}$ (in this case a Multivariate Normal). Furthermore, if it is assumed that the random variable ε^*_d (sum of the two variables τ^*_d and η_d) is distributed as a Gumbel $G(0,\theta_d)$ r.v., the destination choice model is a Multinomial Logit model:

$$p[d]=Pr[V_d + s_d + \varepsilon^*_d \geq V_{d'} + s_{d'} + \varepsilon^*_{d'}]= \frac{exp[(V_d + s_d)/\theta_d]}{\sum_{d'} exp[(V_{d'} + s_{d'})/\theta_d]} \tag{4.2.10}$$

Therefore, the joint destination-path choice probability can be obtained as:

$$p[dk] = p[d](\boldsymbol{V}_d, \boldsymbol{s}_d) \cdot p[k/d](\boldsymbol{V}_{k/d})$$

where the vector $\boldsymbol{V}_{k/d}$ comprises the systematic utilities of all path connecting a certain O-D pair, the vector $\boldsymbol{s}_d$ the EMPU of path choice for all destinations and $\boldsymbol{V}_d$ the systematic utilities of all destinations.

This approach can be extended to all choice dimensions, under the assumption that destination, mode and path systematic utilities and residuals do not depend on the number of trips x, consistently with the specification of the partial share model expressed by (4.2.2):

$$p[x,d,m,k] = p[x](\boldsymbol{V}_x, \boldsymbol{s}_x)\cdot p[d](\boldsymbol{V}_d, \boldsymbol{s}_d)\cdot p[m/d](\boldsymbol{V}_{m/d}, \boldsymbol{s}_{m/d})\cdot p[k/dm](\boldsymbol{V}_{k/dm})$$

where the Expected Maximum Perceived Utility variables are:

$$s_{m/d} = E[\max_{k'}(V_{k'/dm} + \tau_{k'/dm})]$$
$$s_d = E[\max_{m'}(V_{m'/d} + s_{m'/d} + \varepsilon_{m'/d})]$$
$$s_x = E[\max_{d'}(V_{d'} + s_{d'} + \varepsilon_{d'})]$$

and the models expressing the various shares may have any functional form as long as they can be obtained from the assumptions of random utility theory[9].

If each choice dimension is represented by a Logit or a Hierarchical Logit model, then the demand model turns out to be a Hierarchical Logit model with alternatives given by $[x,d,m,k]$, as described in sub-section 3.3.3. In this case, returning to the two-level example, $\tau_{k/d}$ is a Gumbel variable $G(0, \theta_k)$ and so is τ^*_d for the property of stability with respect to maximization of Gumbel variables. Furthermore, the EMPU s_d can be expressed analytically as the product of θ_k and the Logsum variable.

The total four-level partial share model in the case of Hierarchical Logit can be expressed as:

$$p[xdmk] = \frac{exp[V_x/\theta_o + \delta_x Y_x]}{\sum_{x'} exp[V_{x'}/\theta_o + \delta_x Y_{x'}]} \cdot \frac{exp[V_d/\theta_d + \delta_d Y_d]}{\sum_{d'} exp[V_{d'}/\theta_d + \delta_d Y_{d'}]} \cdot \frac{exp[V_{m/d}/\theta_m + \delta_m Y_{m/d}]}{\sum_{m'} exp[V_{m'/d}/\theta_m + \delta_m Y_{m'/d}]} \cdot \frac{exp[V_{k/dm}/\theta_k]}{\sum_{k'} exp[V_{k'/dm}/\theta_k]} \tag{4.2.11}$$

with

$$Y_x = ln\sum_{d'} exp[V_{d'}/\theta_d + \delta_d Y_{d'}] \tag{4.2.12a}$$

$$Y_d = ln\sum_{m'} exp[V_{m'/d}/\theta_m + \delta_m Y_{m'/d}] \quad (4.2.12b)$$

$$Y_{m/d} = ln\sum_{k'} exp[V_{k'/dm}/\theta_k] \quad (4.2.12c)$$

Finally if all δ coefficients are equal to one, i.e. all θ are equal to θ_o, the partial share model turns out to be the factorialization of a Multinomial Logit model with choice alternatives $[x,d,m,k]$.

4.3. Examples of trip demand models

This section describes some models usually adopted within a four-level structure, with the introduction of some possible extensions such as the choice of parking type and location in the context of mode choice model. An example of a whole model system for inter-city travel demand is also described at the end of the section.

4.3.1. Emission or trip frequency models

The emission or trip frequency model estimates the mean number of "relevant" trips $d^i_o[sh]$ undertaken in the period h for the purpose s by the generic user belonging to category i with origin in zone o. The *relevant trips* may be all the trips undertaken for a certain purpose, or the fraction of these which satisfies certain conditions, e.g. car trips or those external to the origin zone. It is clear that in this second case there is a distortion in the interpretation of the model and the variables that appear in it must take into account factors which exclude "non-relevant" trips (in the previous examples, trips with other modes or intra-zonal trips). In this case, the emission model includes elements belonging to other choice dimensions (in the previous examples, mode and destination). For these reasons, in the remainder of this section reference will be made to the case in which the emission model refers to the totality of trips originating in zone o.

The emission models used in applications can be classified in two main categories: behavioral models (or more properly, random utility models) and descriptive models.

To describe the emission model, first define the mean number of trips undertaken by the individual in category i, departing from o, for the purpose s in the period h:

$$m^i[osh] = \Sigma_x xp^i[x/osh]\ (\boldsymbol{SE},\boldsymbol{T}) \quad (4.3.1)$$

The flow of trips from zone o can then be expressed as follows:

$$d^i_o[sh] = n^i[o]\, m^i[osh] \quad (4.3.2)$$

where $n^i[o]$ is the number of users in zone o belonging to category i.

The Binomial and Multinomial Logit are the *random utility models* most frequently used to simulate p^i $[x/osh]$ in 4.3.1. If h is short so that the probability of undertaking more than one "relevant" trip is negligible, a Binomial Logit with two alternatives - undertaking or not undertaking the trip - can be used. Otherwise, a Multinomial Logit gives the probability $p^i[x/osh]$ of undertaking x trips with x equal to $0,1,2,\ldots,n$ or more trips:

$$p^i[x/osh] = \frac{exp(V_x^i / \theta_o)}{\sum_{j=0,\ldots,n} exp(V_j^i / \theta_o)}$$

Systematic utility functions include variables representing the "need" or the "possibility" to carry out activities connected with the purpose examined. These variables may relate either to the family or to the individual. Examples of variables of the former type are income and the number of members of the household, while examples of individual's variables may be occupational status, sex, family role, age, etc. Other variables often used in the systematic utility of trip frequency models relate to the area of origin, and especially its "*accessibility*" with respect to the possible destinations for the trip purpose. Accessibility can be expressed by the Expected Maximum Perceived Utility (EMPU) corresponding to the destination choice model, for example the logsum Y_x given by expression (4.2.12a), in the case of a Logit distribution model. Fig. 4.3.1 gives an example of trip frequency model for the morning peak period in an urban area.

A model of this type should be considered as a tool for quantitative analysis of the determinants of urban mobility[10] rather than an operational tool. Its application for the simulation of travel demand in an entire urban area would require a considerable amount of information. However, this is not necessarily true for all behavioral models and operative trip frequency models are often used for the simulation of large-scale systems; the extra-urban trip frequency models reported in section 4.3.5 are examples of this type of models.

Category index is the simplest *descriptive emission model.* For each category of users, assumed to be homogeneous with respect to a given trip purpose, the average number of trips $m^i[osh]$ for the purpose s in the reference period h is directly estimated. As an example of "category index" models, Fig. 4.3.2 shows the daily home-based work, school and other trip purpose indices obtained as the average of the indices estimated in the mid-80s in five medium-sized Italian towns. Note the different definition of user typology adopted for different trip purposes: the workers in the different economic sectors for Home-Work trips, the students of different levels for Home-School trips, and the family for Home-Other Purpose trips. The main limitation of category index models is that trip frequencies and demand levels are not expressed as functions of socio-economic variables other than those used to define categories; data availability restricts the number of categories to be small.

Category regression models are more sophisticated. These models express the average index $m^i[osh]$ for the generic element of category i for purpose s as a function, typically linear, of variables corresponding to the category and the zone of origin:

$$m^i[osh] = \Sigma_j \beta_j X_{jo}^i \tag{4.3.3}$$

$$V_{TRIP} = \beta_1 CA + \beta_2 WRK + \beta_3 AGE + \beta_4 INL + \beta_5 WMN + \beta_6 ACC$$
$$V_{NOTRIP} = \beta_7 TOP + \beta_8 TOF + \beta_9 NT$$

Typology of variables	Name of variables	
Socio-economic	Car availability	*CA*
	Working status	*WRK*
	Age	*AGE*
	Income level	*INL*
	Woman	*WMN*
Location	Accessibility	*ACC*
Time availability	Nr. of other trips made by the person for other purposes	*TOP*
Individual-family relationships	Nr. of trips of made by other family members for the same purpose	*TOF*
Alternative Specific Attributes (ASA)	NOTRIP	*NT*

CA	dummy variable: 0 = car not available; 1 car available
WRK	dummy variable: 0 = non-worker; 1 = worker
AGE	dummy variable: 0 = ≤ 35 years; 1 = ≥ 35
INL	Income level in 6 points scale: 0 = low income; 5 = high income
WMN	dummy variable: 0 = man, 1 = woman

	No trip			Trip					
	TOP	*TOF*	*NT*	*CA*	*WRK*	*AGE*	*INL*	*WMN*	*ACC*
Shopping	0.55	0.61	1.35	0.24	-2.69	-2.53	0.08	0.60	0.11
t	*5.4*	*3.7*	*5.4*	*1.2*	*-9.7*	*-8.0*	*1.5*	*3.8*	*1.7*
Other purposes	0.22	-1.18	2.66	-	-0.34	-0.34	0.20	0.53	-
t	*2.2*	*-10.9*	*15.3*	-	*-2.0*	*-2.0*	*3.5*	*3.3*	-

	Goodness of fit Statistics		
	ρ^2	*% right*	*LR*
Shopping	0.431	0.847	1904
Other purposes	0.689	0.933	3041

Fig. 4.3.1 Trip frequency model for the morning a.m. peak period.

The attributes X_{jo} are usually the mean values of socio-economic variables such as income, number of cars owned, etc., but they may also include level-of-service attributes such as zonal accessibility, defined by the inclusive variable Y_x, or by some other variable. The name "category regression" is derived from the statistical model, linear regression, used for the specification of variables X_j and the estimation of coefficients β_j. In early applications, model (4.3.3) was specified for traffic zones

(zonal regression). Thus, its explanatory variables represented attributes of an entire zone (e.g., population, number of workplaces, number of shops, etc.).

Recently, more disaggregate categories have been used in these models, typically families and individuals (family or individual regression). The application of model (4.3.3) at a disaggregate level, however, can lead to problems since some variable-coefficient combinations may give negative mobility indices. For this reason, it is better to use Logit or other random utility specifications for disaggregate model representations.

PURPOSE	TYPE OF USER	EMISSION INDEX
H-W	Worker in the Industrial sector	1.024
	Worker in the Services sector	1.084
	Worker in the Private Services sectors	1.245
	Worker in the Public Services sector	0.931
H-Sc	Primary school students	0.84
	Lower secondary school students	0.87
	Upper secondary school students	0.86
	Professional secondary school students	0.88
H-Sndg	Family	0.25
H-Sdg	Family	0.11
H-Ps	Family	0.16
H-Sr	Family	0.27
H-Acc	Family	0.11
H-oth	Family	0.13

Trip identification	
H-W	*Home-Work*
H-Sc	*Home-School*
H-Sndg	*Shopping for non-durable goods*
H-Sdg	*Shopping for durable goods*
H-Ps	*Personal services*
H-Sr	*Social-recreational*
H-Acc	*Accompaniment of others*
H-Oth	*Other purposes*

Fig. 4.3.2 Daily urban trip emissions indices.

Clearly, random utility models (4.3.1), or family or individual regression models (4.3.3) require more information[11] than the category index model (4.3.2). The latter however has the shortcoming of not being elastic with respect to variations of variables other than those used to define the category.

Finally, both for random utility models and descriptive models, the emission model for purpose s in the reference period h is sometimes decomposed into two models: an emission model $m^i[o,s]$ over a longer period, e.g. the whole day, and a time-distribution model:

$$m^i[osh] = \sum_x p^i[x/os]\ (\boldsymbol{SE},\boldsymbol{T}) \cdot \sum_{y_h=i}^{x} y_h p^i[y_h/osx]\ (\boldsymbol{SE},\boldsymbol{T})$$

The fraction of the x trips undertaken during the time period h, $p^i[y_h/osx]$, can be obtained with a random utility model simulating the allocation of trips (or activities) among available time periods. This case can be handled as yet another stage in the traditional structure of the partial share model, which was introduced in equation (4.2.3).

4.3.2. Distribution models

Distribution models express the percentage (probability) $p^i[d/osh]$ of trips undertaken by users of category i going to destination d, given departure from zone o, purpose s, and period h. For simplicity of notation, the category index will be omitted and it will be assumed, as in expression (4.2.2), that the number x of trips does not affect the distribution among destinations.

Typically, distribution models have a Multinomial Logit structure:

$$p[d/osh](\boldsymbol{SE},\boldsymbol{T}) = \frac{exp(V_d/\theta_d)}{\sum_{d'} exp(V_{d'}/\theta_d)} \tag{4.3.4}$$

The systematic utility V_d is a linear combination of the attributes of possible destinations in relation to the zone of origin o:

$$V_d = \sum_j \beta_j X_{jd} \tag{4.3.5}$$

In general, the attributes in function (4.3.5) can be divided into two categories: attributes representing the "attractiveness" of zone d, i.e. the convenience of carrying out activity s in d, and "cost" attributes which represent the inconvenience of undertaking a trip from o to d. Distribution models can be interpreted and specified following either a behavioral or a descriptive approach with various specifications and interpretations of the attributes.

According to the *behavioral interpretation*, the distribution model simulates the choice of a destination among possible alternatives. It should be noted that typically the destination chosen for carrying out an activity is not a traffic zone but one (or more) "elementary" destination (such as a shop, an office, etc.) within it. The traffic zone d is therefore a "compound" alternative composed of the aggregation of M_d elementary alternatives.

If a positive covariance between the perceived utilities of elementary destinations belonging to a same zone d is assumed, the utility of each elementary alternative r of a zone d, U_{rd}, can be expressed as:

$$U_{rd} = U_d + U_{r/d} = V_d + \eta_d + V_{r/d} + \tau_{r/d} \tag{4.3.6}$$

that implies:

$$Cov[U_{r/d}, U_{s/d}] = Var[\eta_d] \qquad \forall r, s \in d; \ \forall d \tag{4.3.7}$$

while the utility of each compound alternative d ,U^*_d, becomes:

$$U^*_d = max_{r'}\{U_{r'd}\} = U_d + max_{r'}\{U_{r'/d}\} = V_d + s_d + \eta_d + \tau^*_d = V_d + s_d + \varepsilon^*_d$$

where s_d is the EMPU of elementary destinations of zone d:

$$s_d = \mathrm{E}[max_{r'}\{U_{r'/d}\}]$$
$$\tau^*_d = [max_{r'}\{U_{r'/d}\}] - \mathrm{E}[max_{r'}\{U_{r'/d}\}]$$

If the *r.v.* $\tau_{r/d}$ are assumed i.i.d. Gumbel, $\tau_{r/d} \sim G(0, \theta_r)$, it results:

$$s_d = \theta_r Y_d = \theta_r ln \sum_{r'=1,...M_d} exp(V_{r'/d} / \theta_r) \tag{4.3.8}$$
$$\tau^*_d \sim G(0, \theta_r)$$

and if also the *r.v.* ε^*_d are assumed i.i.d. Gumbel, $\varepsilon^*_d \sim G(0, \theta_d)$ the choice probability of a zone d becomes:

$$p[d] = \frac{exp(V_d / \theta_d + \delta_d Y_d)}{\sum_{d'} exp(V_{d'} / \theta_d + \delta_d Y_{d'})} \tag{4.3.9}$$

where:

$$\delta_d = \theta_r / \theta_d \leq 1$$

and the (4.3.7) becomes:

$$Cov(U_{r/d}, U_{s/d}) = \pi^2(\theta_d^2 - \theta_r^2)/6 = \pi^2 \theta_d^2 (1 - \delta_d^2)/6 \qquad \forall r, s \in d; \ \forall d$$

Alternatively a null covariance between the perceived utilities of elementary destinations belonging to a same zone d can be assumed posing $\eta_d = 0$ in equation (4.3.6), i.e. $\theta_r = \theta_d$. In this case the utility of each compound alternative d, U^*_d, becomes:

$$U^*_d = V_d + max_{r'}\{U_{r'/d}\} = V_d + s_d + \tau^*_d$$

and the choice probability of a zone d becomes:

$$p[d] = \frac{exp(V_d / \theta_d + Y_d)}{\sum_{d'} exp(V_{d'} / \theta_d + Y_{d'})} \tag{4.3.10}$$

It results generally hard the evaluation of the EMPU variable (4.3.8) (and thus also that of the logsum variable Y_d in (4.3.9) and (4.3.10)) since it is generally hard the evaluation of the V_{rd} terms. To overcome this problem, this variable, through some analytic manipulation, can be expressed as a function of the mean systematic utility $\bar{V}_d$ of the elementary destinations:

$$\bar{V}_d = \frac{1}{M_d} \sum_{r'=1,\ldots,M_d} V_{r'/d}$$

as well as the number of elementary alternatives M_d and a heterogeneity term taking into account the variability of the elementary utilities V_{rd} compared with the mean value $\bar{V}_d$:

$$s_d = \theta_r Y_d = \bar{V}_d + \theta_r ln M_d + \theta_r ln\left[\frac{1}{M_d} \sum_{r'=1,\ldots,M_d} exp\left[\left(V_{r'/d} - \bar{V}_d\right)/\theta_r\right]\right] \qquad (4.3.11a)$$

In (4.3.11a), the second and the third terms are non-negative ($M_d \geq 1$, $\Sigma_{r'} exp(.) \geq M_d$) and thus, consistently with the properties of the EMPU variable described in section 3.5, s_d is larger than the mean of the systematic utility of the elementary destinations.

Also, if all the elementary destinations in d had the same systematic utility (e.g., the generalized trip cost and an attractivity value equal for each elementary destination), the heterogeneity term would be equal to zero. In this case, s_d would be equal to the sum of the constant utility of each elementary destination, and a positive "size" variable ($ln\ M_d$). The latter increases with the number of elementary alternatives included in d.

In applications $\bar{V}_d$ is generally simulated through attributes of the zone, X_{jd}. Moreover, an estimate of the heterogeneity term is unlikely to be available and for this reason it is usually omitted or replaced by a "proxy" (e.g. the variance of the number of workers in each elementary destination). In this case, the proxy variables are included in the attributes of the zone and thus it results:

$$s_d = \sum_j \beta_j X_{jd} + \theta_r ln M_d \qquad (4.3.11b)$$

For certain trip purposes it is possible to assume that the number of elementary destinations, M_d, within each traffic zone d can be measured (e.g., the number of shops for the shopping purpose). However, for several trip purposes the precise identification of the elementary destinations and their number is not possible. In this case the size variable $ln\ M_d$ may be replaced by a "*size function*" expressing the actual (unknown) number of elementary destinations as a function of other variables

(such as population, employment in different sectors, number of firms of different types, etc.):

$$M_d = \sum_k \beta_k Z_{kd}$$

In this case it can be shown that the coefficients β_k of the size function can all be identified except one, which is arbitrarily set equal to one (see Chapter 8) (since we are actually calibrating the products $\theta_r \beta_k$) and thus equation (4.3.11b) becomes:

$$s_d = \sum_j \beta_j X_{jd} + \theta_r ln\left(Z_{1d} + \sum_{k=2,\ldots,K} \beta_k Z_{kd} \right) \qquad (4.3.11c)$$

It should also be noted that, given the behavioral interpretation of the distribution model, one of the zone attributes, X_{jd}, should be the inclusive variable Y'_d representing lower choice dimensions, typically mode choice, expressed by (4.2.12b) in the case of Logit model.

In this case, equation (4.3.11c) becomes:

$$s_d = \theta_m Y'_d + \sum_{j=1,\ldots,J} \beta_j X_{jd} + \theta_r ln\left(Z_{1d} + \sum_{k=2,\ldots,K} \beta_k Z_{kd} \right) \qquad (4.3.11d)$$

with

$$\theta_m \leq \theta_r$$

Some examples of destination choice models with size functions are reported in section 4.3.5.

Descriptive distribution models generally have a simpler functional form than behavioral models. However, in many cases descriptive distribution models can be recast in a Multinomial Logit structure (4.3.4). The variables included in descriptive distribution models can be divided into two groups: attributes of the activity system located in zone d, or attractivity attributes, and cost or separation attributes between zones o and d. *Attractivity attributes* are variables or functions similar to those described for behavioral models, although their interpretation may be different. Examples of such variables are total (or sectorial) employment for home-work trips, the number of students attending schools for home-school trips, retail employment for home-shopping trips, etc. *Cost attributes* are variables measuring the generalized cost of a trip from o to d; their coefficients β_k are therefore negative. Several descriptive cost attributes have been proposed in the literature, from the crow-flight distance between zone centroids, to generalized cost variables including various components (on foot, on board times, monetary cost, etc.) for the different transport modes available. The most elementary descriptive distribution model has the form:

$$p[\beta_1 A_{d'} - \beta_2 C_{od'}] = \frac{exp(\beta_1 A_d - \beta_2 C_{od})}{\sum_{d'} exp(\beta_1 A_{d'} - \beta_2 C_{od'})} \qquad (4.3.12)$$

where A_d and C_{od} are respectively an attraction variable and a cost variable. In applications, special forms of the model (4.3.12) called *simply constrained gravitational models* are sometimes used[12]. A model of this type is obtained by taking the natural logarithm of the attraction variable. Substituting $ln(A'_d)$ for A_d in expression (4.3.12) yields:

$$p[d/osh] = \frac{A'^{\beta_1}_d \, e^{-\beta_2 C_{od}}}{\sum_{d'} A'^{\beta_1}_{d'} \, e^{-\beta_2 C_{od'}}} \tag{4.3.13}$$

If parameter β_1 is equal to one, specification (4.3.13) is invariant with respect to the aggregation or disaggregation of traffic zones, given equal "distance" from the origin. In this case, the probability of going to a zone d obtained by aggregating two smaller zones d_1 and d_2, is equal to the sum of the probabilities of d_1 and d_2. In fact, if the cost is constant ($C_{od} = C_{od_1} = C_{od_2}$) and the attraction variable associated with d is the sum of those for zones d_1 and d_2 ($A_d = A_{d_1} + A_{d_2}$) it follows:

$$p[d/osh] = \frac{A_d \, e^{-\beta_2 C_{od}}}{A_d \, e^{-\beta_2 C_{od}} + \sum_{d' \neq d} A_{d'} e^{-\beta_2 C_{od'}}} = \frac{A_{d_1} \, e^{-\beta_2 C_{od_1}}}{A_{d_1} e^{-\beta_2 C_{od_1}} + A_{d_2} \, e^{-\beta_2 C_{od_2}} + \sum_{d'} A_{d'} \, e^{-\beta_2 C_{od'}}} +$$

$$+ \frac{A_{d_2} \, e^{-\beta_2 C_{od_2}}}{A_{d_1} e^{-\beta_2 C_{od_1}} + A_{d_2} \, e^{-\beta_2 C_{od_2}} + \sum_{d'} A_{d'} \, e^{-\beta_2 C_{od'}}} = p[d_1 / osh] + p[d_2 / osh]$$

The property of invariance with respect to zone aggregation is convenient in applications since it eliminates the influence of the level of spatial disaggregation adopted.

If the logarithmic transformation of the cost variable C_{od} ($ln(C_{od})$) is carried out, model (4.3.12) becomes:

$$p[d/osh] = \frac{A_d^{\beta_1} \cdot C_{od}^{-\beta_2}}{\sum_{d'} A_{d'}^{\beta_1} \cdot C_{od'}^{-\beta_2}} \tag{4.3.14}$$

4.3.3. Mode choice models

Mode choice models simulate the fraction (probability) $p^i[m/oshd]$ of trips of category i users using mode m, from zone o to zone d for trip purpose s in time period h. Mode choice is a typical example of a travel choice that can be modified for different journeys in which performance or level-of-service attributes have considerable influence. It was not simply by chance that the first random utility models were formulated with reference to transport mode choice.

The identification of *relevant alternatives* (the *choice set*) depends on the transport system under study. For example, in an urban system, modes such as "walking" or "bicycle" are typically considered to be choice alternatives while, for obvious reasons, they are not included for interurban systems. In some cases "mixed" modes, i.e. combinations of different modes such as car + train and car + bus, or different services of the same transport mode (e.g., Intercity, Regional and Night for the railway mode) are included as choice alternatives. For modal choice models, the definition of the choice set of each decision-maker (described in Chapter 3) is particularly important. In fact, not all transport modes are available for all trips, either because of an objective impossibility (e.g., the personal car is not available to a user without driving license) or because it is not perceived as an alternative for a particular trip (e.g. motorized modes are not considered for very short trips).

Mode availability has been handled in mode choice models using the different approaches described in section 3.4, usually with a combination of several heuristic methods. "Objective" non-availability is usually dealt with by excluding the alternatives from the decision-maker's (or category of decision-makers) choice set; while "contingent" non-availability or non-perception is generally simulated by including "availability/perception variables" in the systematic utility specification. The attributes of car, bicycle and motorcycle availability in the specification described in Fig. 4.3.3 should be interpreted in this way. Recently, IAP models that implicitly simulate the probability of an alternative being available/perceived (as described in section 3.4) have been applied to mode choice.

Attributes in the systematic utility functions of mode choice models are usually level-of-service and socio-economic attributes. As stated in Chapter 2, *level-of-service* or *performance attributes* describe the characteristics of the service offered by the specific mode. Examples are travel time (possibly decomposed into access/egress time, waiting time, on-board time, etc.), monetary cost, regularity of the service, number of transfers and so on. These attributes have negative coefficients since they usually represent disutilities for the user. In addition to level-of-service attributes, it is possible to include Alternative Specific Attributes (ASA) or *modal preference*, variables which account for qualitative characteristics of each mode (e.g., the privacy of the car) or those not explicitly included in the attributes (e.g. service regularity for metro systems). In Chapter 3 it was shown that in order to estimate ASA coefficients (ASC) in additive random utility models, ASA variables may be included in the systematic utility expressions for all but one alternative. Thus, ASA variables represent the relative preference of each mode with respect to a reference alternative that remains unexplained by other attributes. On the basis of this interpretation, the ASC might have a positive or a negative sign.

The ratios between coefficients of level-of-service attributes, also called *reciprocal substitution rates*, are considered to be important. Among these, in the context of project evaluation discussed in Chapter 10, the substitution rates with monetary cost are particularly relevant, i.e. the monetary value of level-of-service attributes. If β_t and β_c are respectively the coefficients of travel time and monetary cost, the perceived Value Of Time (VOT) implicit in modal choice behavior will be:

$$\text{VOT} = \frac{\beta_t}{\beta_c} \frac{[h^{-1}]}{[mon.unit^{-1}]} = [mon.unit/h] \tag{4.3.15}$$

Level-of-service attributes, and in particular times, monetary costs, etc., should take into account alternatives on the "lower" choice dimension, in this case path choice. Thus, level-of-service attributes should refer to the different routes that the user can take on the network of each mode. This is done by using the EMPU of route choice, which, in Logit or Hierarchical Logit models, is the logsum variable $Y_{m/d}$. Sometimes, for the sake of simplicity, attributes are calculated only for the "minimum" cost route though this introduces a theoretical inconsistency if route choice is not simulated with the deterministic utility (minimum cost) model described in the next section.

Socio-economic attributes are characteristics of the decision-maker or his/her family. Typical examples are gender, age, family income, car ownership or availability (number of cars owned by the family or the ratio between the cars owned and number of driving licenses), etc. Since socio-economic attributes don't depend on the alternative, for additive models with linear in coefficients systematic utilities, they cannot be included in the systematic utilities of all alternatives.

Finally, in more sophisticated specifications some attributes may depend jointly on service and user characteristics. For example, monetary cost can be divided by user income, or differentiated by income level with different coefficients. In both cases the value of time (VOT) is differentiated on the basis of income, and is usually higher for users with higher income.

With respect to the functional form, Multinomial Logit modal choice models are often used:

$$p^i[m/oshd] = \frac{exp(V^i{}_{m/oshd})}{\sum_{m'} exp(V^i{}_{m'/oshd})} \tag{4.3.16}$$

Fig. 4.3.3 shows the alternatives, attributes and coefficients of a Logit mode choice model for commuting trips in an Italian medium-sized city. Other examples of *MNL* modal choice models are reported in section 4.3.5 and in Chapter 8 on the estimation of transportation demand.

Hierarchical Logit specifications are also being increasingly used. These models assume different levels of correlation between the perceived utilities of different mode groups, for example individual modes and public modes, and/or between different services of the same mode. Fig. 3.3.7 shows a possible choice tree and the associated correlation structure for an inter-city mode choice model. As another example, a Hierarchical Logit mode choice model could be used to simulate jointly mode choice and parking in urban areas.

WALKING		
$T_{walking}$	Time (hr.)	-6.8237
BICYCLE		
T_{bk}	Time (hr.)	-8.2718
Nbcl/Nad	Number of bicycles owned in family per adult	0.6646
Bcl	Alternative Specific Attribute	-1.5818
MOTORCYCLE		
T_{mbk}	Time (hr.)	-8.2718
Age	Age variable (1 if ≤ 35 years 0 otherwise)	0.6863
Nmbk/Nad	Number of scooters and motorbikes owned in family per adult	1.8572
Mbk	Alternative Specific Attribute	-2.3789
CAR		
T_{car}	Time (hr.)	-1.6142
Mc_{car}	Monetary cost (€)	-0.3338
Park	Parking (1 for priced parking destinations, 0 otherwise)	-1.1469
Hfam	Position in the family (1 if head of family, 0 otherwise)	0.4931
Ncar/Nad	Number of cars owned in family per adult	0.4014
Car	Alternative Specific Attribute	-1.7103
BUS		
T_{bus}	Total travel time (hr.)	-1.6142
Mc_{bus}	Monetary cost (€)	-0.3338
Ntrn	Number of transfers	-01772
Bus	Alternative Specific Attribute	-1.7827
$lnL(\beta_{ML})$		-475
$lnL(0)$		-697
ρ^2		0.317
% right		0.651

Fig. 4.3.3 Alternatives, attributes and coefficients of a MNL mode choice model for urban commuting trips.

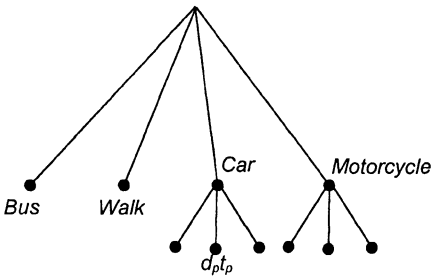
In some applications to urban areas, the specification of the systematic utility of the car mode includes level-of-service attributes related to parking, such as the time spent looking for a free parking space, walking distance, fare. In the most general case where several locations and types of parking are available, individual modes, e.g. auto, are simulated as groups of alternatives each corresponding to a specific parking location (d_p) and parking type (t_p) with the given mode, see Fig. 4.3.4.

The lower-level Multinomial Logit model for parking choice can be specified as follows:

$$p^i[d_p t_p / oshda] = \frac{exp(V^i_{d_p t_p})}{\sum_{d'_p t'_p} exp(V^i_{d'_p t'_p})}$$

with

$$V^i{}_{d_p t_p} = \beta_{ts} Tsr_{d_p t_p} + \beta_c Mc^i_{d_p t_p} + \beta_{tw} Twl_{d_p / d}$$



$$V_{car} = \beta_{tb} \cdot T_{car} + \delta_p \cdot Y_p + \beta_c \cdot Mc_{car} + \beta_{Car} \cdot Car$$
$$V_{mbk} = \beta_{tb} \cdot T_{mbk} + \delta_p \cdot Y_p + \beta_c \cdot Mc_{mbk} + \beta_{Age} \cdot Age + \beta_{Mbk}\, Mbk$$
$$V_{bus} = \beta_{tb} \cdot T_b + \beta_{tw} \cdot tw_b + \beta_c \cdot Mc_b$$
$$V_{walk} = \beta_{twalk} \cdot Twalk + \beta_{Walk} \cdot Walk$$

with

T_{car} = Car travel time [h]
T_{mbk} = Motorbike travel time [h]
Tb_b = Bus on board time [h]
Tw_b = Bus waiting time [h]
Twl = Walking time [h]
Age = Dummy variable of value 1 if age is < 35 years, 0 otherwise
$Car, Mbk, Walk$ = Mode Specific Attributes

Mc_{car} = Monetary cost Car [€]
Mc_{mbk} = Monetary cost Motorbyke [€]
Mc_b = Monetary cost Bus [€]

$$Yp = ln \sum_{d'_p t'_p} exp(V^i_{d'_p t'_p})$$

d_p= parking destination zone

t_p=type of parking:

- free limited duration
- toll on street
- tool off street
- illegal

$$V_{d_p t_p} = \beta_{ts} Tsr_{d_p t_p} + \beta_c Mc^i_{d_p t_p} + \beta_{tw} Twl_{d_p / d}$$

With

Tsr = Average time spent finding a parking space [h]
Mc = Parking monetary cost [€]
Twl = Walking time from parking location to destination [h]

Model of parking choice

Tsr	*Mc*	*Twl*
-18.168	-3.358	-19.386

Model of modal choice

T	*Tw*	*Twl*	*Mc*	Y_p	*Age*	*Car*	*Mbk*	*Walk*
-1.961	-4.902	-4.314	-0.550	0.199	2.331	0.921	-1.631	3.127

Fig. 4.3.4 A Hierarchical Logit model of mode and parking choice in an urban area.

where the variables indicate:

d_p, t_p parking location (zone) and type (free on street, toll on-street, toll off-street, illegal etc.);

$Tsr_{d_p t_p}$ average search time to find a parking space of type t_p in the zone d_p;

$Mc^i_{d_p t_p}$ monetary cost (price or expected fine) of the alternative depending on the category of users i (e.g., related to the parking duration);

$Twl_{d_p/d}$ time on foot needed to reach final destination d from location d_p.

In this case, the logsum inclusive variable Y^i_pcan be expressed as:

$$Y^i_p = ln \sum_{d'_p t'_p} exp(V^i_{d'_p t'_p})$$

and included in the car systematic utility of the *MNL* model simulating choice among modes.

An example of Hierarchical Logit mode and parking choice model in an urban area is reported in Fig. 4.3.4.

4.3.4. Path choice models

The path choice model provides the fraction (probability) $p^i[k/oshdm]$ of the trips undertaken by users of category i, using route k on mode m from o to d for trip purpose s in the time period h. Path choice models used in practice are all behavioral and the relevant attributes are mostly performance or level-of-service variables obtained from the network supply models described in Chapter 2.

Path choice behavior and the models representing it depend on the type of service offered by the different transport modes. In particular, the case in which the whole path is chosen before starting the trip *(pre-trip choice)* can be distinguished from that in which the route is chosen in two phases and it is completely defined only during the trip itself (*pre-trip/en route mixed choice*). Pre-trip choice behavior is usually assumed to simulate route choice for continuous service systems; the typical examples are road networks for individual modes like car, motorcycle, etc. Pre-trip choice behavior is also assumed for scheduled transport services with sufficiently low frequency and high regularity under the assumption that the user knows the service timetable and makes his/her decisions before beginning the trip (see section 6.5.1). On the other hand, pre-trip/en route mixed behavior is usually assumed for scheduled transport systems with high frequency and/or low regularity, as is the case for urban transit systems (see section 5.5).

As for all behavioral models, the complete specification of a path choice model can be decomposed in three phases: definition of the alternatives; identification of the set of possible alternatives (choice set) and definition of the choice model. The first two phases are particularly important for path choice.

In the following, behavioral assumptions and choice models will be described separately for pre-trip and mixed path choice behavior with regard to road continuous systems and to transit networks with high frequency/low regularity. Path choice models for low frequency/high regularity scheduled services will be described in Chapter 6.

4.3.4.1. Path choice models for road systems

Definition of choice alternatives. The hypothesis usually accepted for road systems is that the user, before undertaking the trip, chooses a sequence of road segments to

follow, or the phases of the trip, which can be represented as a path[13]. This is a sequence of nodes and links on the graph representing the road system as described in Chapter 2. Only elementary (loop-less) paths are considered, and thus their number is finite.

Identification of the choice set. The definition of the paths considered as choice alternatives, i.e. the definition of the choice set, is particularly relevant since the topological complexity of the network could generate an unrealistic high number of routes connecting a single O-D pair. The set of feasible routes K_{odm} connecting the centroid pair *od* on the mode network *m* should be defined through an explicit choice set model, as described in section 3.4. In practice, however, heuristic approaches of two types are used.

The *exhaustive approach* considers all elementary paths on the network. This approach may generate many routes that share many links; thus, these routes are correlated in their perceived (dis)utilities. Furthermore, given the computational complexity of explicitly enumerating all the routes in a network, this operation is usually carried out implicitly (*implicit path enumeration*) by using implicit algorithms for the calculation of path choice probabilities and flow assignment, as will be described in Chapter 7.

The *selective approach*, on the other hand, identifies only some elementary paths on the basis of heuristic behavioral rules. For example, a route may not include more than one entrance and one exit from the same motorway, may not go "further away" from the destination, may not have a generalized cost exceeding the minimum cost by a given amount, etc. Various criteria for the selection of feasible routes have been proposed in the literature. They correspond to different application contexts (urban/extra-urban networks) and to different algorithms for generating the routes and calculating choice probabilities and link flows. Some examples of selection criteria are given in Fig. 4.3.5 where $Z_{o,i}$ and $Z_{o,j}$ represent the minimum cost to reach node *i* and node *j* from the origin *o* (see Chapter 7).

In general the selective approach requires *explicit path enumeration* between each O-D pair, and usually a combination of criteria is adopted. Chapter 7 describes some algorithms for path enumeration, and Fig. 4.3.6 depicts an example of the complete set of elementary paths and a selective set for an origin-destination pair. For more sophisticated feasible path generation models, the criteria to be used must be "calibrated" like other parameters in the model. Calibration can be carried out by comparing, the paths generated by the model with the paths perceived (or at least with those actually chosen) by a sample of users, in order to maximize the coverage of the latter with the former.

SELECTION CRITERIA	SPECIFICATION
Topological	A path is feasible (Dial efficient) if each link "goes away" from the origin and/or "move towards" the destination, see section 7.3.1a $k \in K_{od}$ if $Z_{o,i} < Z_{o,j}\ \forall (i,j) \in k$
Comparison of costs	Paths with a generalized cost not exceeding by more than α the minimum cost $k \in K_{od}$ if $g_k \leq (1+\alpha)\ g_{min}$
Progressive	The first n minimum generalized cost paths
Multi-attribute	Minimum paths with respect to various attributes (usually the relevant performance variables such as: travel time, monetary cost, motorway distance, etc.)
Behavioral	Paths excluding behaviorally unrealistic link sequences (e.g. repeated entrances and exits for the same motorway)
Distinctive	Paths overlapping for no more than a given percentage of their length

Fig. 4.3.5 Criteria for path feasibility on road networks.

Some experimental results suggest that a good level of coverage of the routes used by a sample of users, at least for extra-urban networks, can be achieved by generating the first n paths for some criterion (e.g. minimum time, minimum monetary cost, maximum motorway use, etc.).

The selective approach guarantees better control of the "feasibility" of the generated routes while allowing the use of performance attributes that are not additive over links as will be seen later[14]. These advantages are obtained at the expense of greater computational complexity.

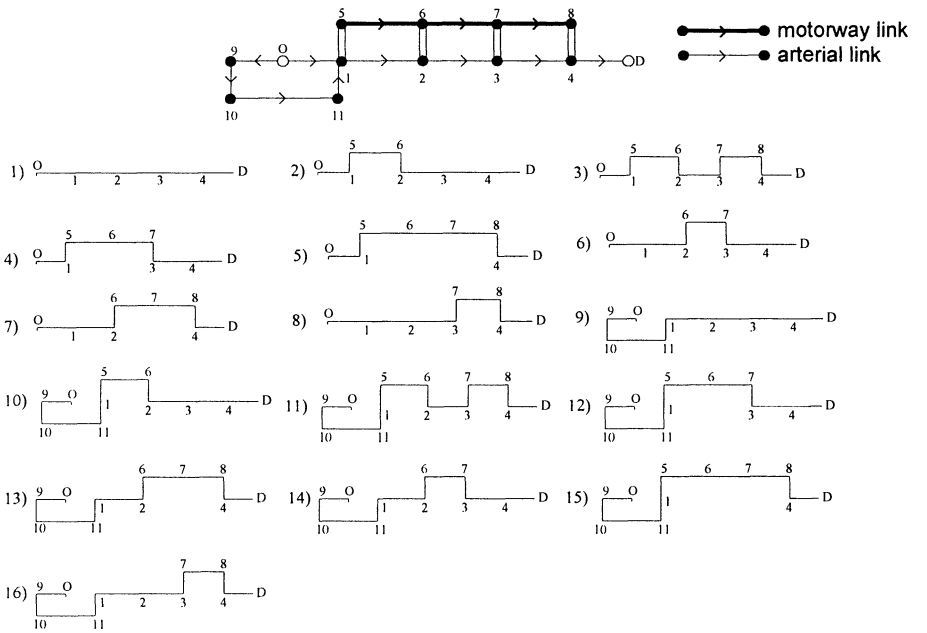
Conversely, implicit path enumeration methods are computationally more efficient and are used in assignment models in commercial software. However, it should be emphasized that no systematic analysis of the computational complexity and memory requirements of the two assignment algorithms exists and that the literature seems to suggest a tendency towards explicit path enumeration models in applications (see Chapter 7).

Specification of the choice model. The specification of the path choice model requires, as usual, definition of the attributes in the systematic utility function and of the joint probability distribution of random residuals, i.e. the choice probability functional form. It is usually assumed that the variables influencing path choice are performance attributes with negative coefficients[15], e.g. travel times, monetary cost, distance, etc. Thus it follows that:

$$U_k = V_k + \varepsilon_k \qquad \forall k \in K_{odm} \tag{4.3.17a}$$

$$V_k = -g_k \tag{4.3.17b}$$

where g_k is the average generalized cost of path k expressed in utility units and K_{odm} is the set of paths connecting the pair o d via mode m. Systematic utility and average cost should be differentiated by user category, V^i_k and g^i_k, although in what follows the superscript i will be omitted for simplicity of presentation.



Paths 1-16: exhaustive set of elementary paths
Paths 1-2 + 4-8: selective set of paths on the basis of behavioral (eliminating paths 3 and 11) and topologic (eliminating paths 9 to 16 going "further away" from the destination criteria)

Fig. 4.3.6 Examples of exhaustive and selective set of feasible paths.

In section 2.2.2 it was stated that the average path cost is usually a linear combination[16] of performance attributes with coefficients estimated with a path choice model:

$$g_k = \Sigma_n \beta_n z_{nk} \tag{4.3.18a}$$

If each attributes z_{nk} can be obtained as a sum of the corresponding link variables r_{nl}, the path cost g_k will be purely additive:

$$c_l = \Sigma_n \beta_n r_{nl} \qquad \boldsymbol{g}^{ADD} = \Sigma_n \beta_n \boldsymbol{z}_n = \Sigma_n \beta_n \Delta^{\mathrm{T}} \boldsymbol{r}_n = \Delta^{\mathrm{T}} \boldsymbol{c} \tag{4.3.18b}$$

where δ_{lk} are the (0/1) elements of the link-path incidence matrix Δ and c_l is the average cost of link l introduced in Chapter 2.

In some cases, the average cost might include some variables that cannot be obtained as the sum of link variables (*non-additive cost* $g^{NA}{}_k$). This occurs, for example, if the monetary cost depends non-linearly on the path length, or if there is a dummy variable for minimum travel time or maximum motorway length paths. In the most general case, the expression (4.3.17b) therefore becomes:

$$V_k = - g_k^{ADD} - g_k^{NA}$$

Generally, a non-additive path cost variable requires explicit path enumeration.

Fig. 4.3.7 shows some examples of systematic utility specification for path choice models in urban and extra-urban road systems environments.

PATH CHOICE MODEL FOR URBAN ROAD NETWORKS

$$V_k = \beta_1 TTP_k + \beta_2 TTS_k + \beta_3 L_k + \beta_4 NTS_k + \beta_5 NLT_k + \beta_6 MTW_k$$

TTP = Travel time on primary roads [*h*]
TTS = Travel time on secondary roads [*h*]
L = Total length [*km*]
NTS = Number of traffic-signal intersections on the path
NLT = Number of left turns
MTW = Dummy variable for the maximum motorway path

	TTP	*TTS*	*L*	*NTS*	*NLT*	*MTW*	ρ^2	*%right*	*L ratio*
	- 16.462	-61.257	-9.601	-0.209	-2.296	3.158	0.403	0.532	844.344
t	*-7.514*	*-16.445*	*-1.224*	*-1.143*	*-3.978*	*2.678*			

PATH CHOICE MODEL FOR HEAVY VEHICLES IN EXTRA-URBAN ROAD NETWORKS

$$V_k = \beta_1 TT_k + \beta_2 Mc_k + \beta_3 ML_k + \beta_4 MinT_k + \beta_5 MaxM_k + \beta_6 HVP_k + \beta_7 CF_k$$

TT = Travel time [*h*]
Mc = Monetary cost [€]
ML = Total motorway length [*km*]
MinT = Dummy variable for minimum time path (0/1)
MaxM = Dummy variable for maximum motorway use path (0/1)
HVP = Dummy variable for minimum time path for perishable and/or high value goods (0/1)
CF = Path commonality factor
VOT = Value of time [€/h]

	TT	*Mc*	*ML*	*MinT*	*MaxM*	*HVP*	*CF*	VOT	ρ^2	*LRratio*
	-4.525	-0.0165	0.013				-0.9524	68.5605	0.176	-2440
t	*-19.3*	*-6.7*	*12.3*				*-12.9*			
	-3.110	-0.0155	0.012			1.785	-0.839	50.1515	0.250	-2222
t	*-14.2*	*-6.1*	*11*			*20.8*	*-11.6*			
	-5.440	-0.018	0.012	2.292	2.585		-1.296	75.5555	0.306	-2055
t	*-20.5*	*-6.9*	*10.5*	*19.9*	*20.1*		*-15.7*			
	-3.650	-0.015	0.009	3.370	3.702	3.788	-1.205	60.8335	0.450	-1630
t	*-14.1*	*-5.6*	*7.5*	*21.6*	22	*22.1*	*-14*			

Fig. 4.3.7 Examples of Multinomial Logit path choice models in urban and extra-urban road networks.

The probability of choosing path k can be obtained with any random utility model whose depending on the distribution of random residuals ε_k in (4.3.17).

The first model used to simulate path choice is the *deterministic utility* model, which is a special case of a random utility model in which the variance of the residuals ε_k is assumed equal to zero:

$$U_k = V_k = -g_k$$

In this case, path k can be used only if its cost g_k is the least among those of alternative paths:

$$p[k / osdm] > 0 \Rightarrow g_k \leq g_h \quad \forall h \neq k \quad h,k \in K_{odm} \tag{4.3.19}$$

As already noted in section 3.5, the deterministic utility model does not provide a unique path choice probability vector, except in the case in which there is a unique minimum cost path. In this case:

$$\begin{aligned} p[k / osdm] &= 1 \text{ if } g_k < g_h \quad \forall h \neq k \quad h,k \in K_{odm} \\ &= 0 \text{ otherwise} \end{aligned} \tag{4.3.20}$$

The deterministic choice model, though less realistic than probabilistic models, is still used for computational reasons in the case of very congested networks with implicit path enumeration. In fact, in those cases, it gives results that are largely comparable with those obtained by using probabilistic models, as will be seen in section 5.4.3.

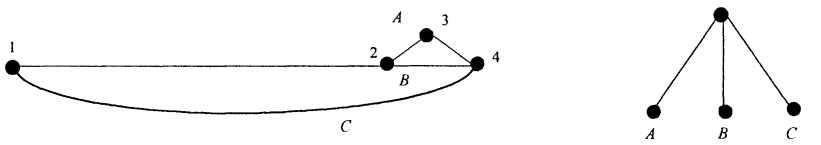
The probabilistic choice models generally used to calculate path choice probability are Logit and Probit. In this case, the *Multinomial Logit* model takes the form:

$$p[k / oshdm] = \frac{exp(-g_k / \theta)}{\sum_{h \in K_{odm}} exp(-g_h / \theta)} \tag{4.3.21}$$

The Multinomial Logit model results from the hypothesis that the random residuals ε_k are i.i.d. Gumbel variables of parameter θ, with θ proportional to the standard deviation of random residuals ε_k. As will be seen in Chapter 8, the parameter θ cannot be estimated separately for linear utility functions of the type (4.3.18a) and is therefore included in the coefficients β_h. The urban and extra-urban path choice models described in Fig. 4.3.7 have a Multinomial Logit specification.

The assumption of identically and independently distributed random residuals that underlies the Logit model and its property of Independence of Irrelevant Alternatives (see section 3.3.1) are unrealistic when alternative paths share several links. In this case, it may be conjectured that the perceived costs of heavily overlapping paths are highly correlated, giving rise to choice probabilities smaller

than those of other paths with the same average costs but with no overlapping. In the extreme case of two practically coincident paths, the *MNL* model gives unrealistically large choice probabilities as is shown in Fig. 4.3.8. Thus, the Multinomial Logit model should be used with an explicit path enumeration eliminating highly overlapping paths to reduce the effects of the IIA property.



Path	Links	
A	(1.2) (2.3) (3.4)	$g_A=g_B=g_C=g$
B	(1.2) (2.4)	$p[A]=p[B]=p[C]=0{,}333$
C	(1.4)	$p[A]+p[B]=0{,}66$

Fig. 4.3.8 Application of a Logit model in the case of highly overlapping paths.

Alternatively, if it is assumed that the residuals ε_k are distributed according to a Multivariate Normal variable, the choice model has the *Probit* form. The most widely used specification assumes that the variance of the random residuals is proportional to an additive path cost attribute, z_k, and that the covariance of the residuals of two paths is proportional to the cost attribute of the links shared by the two paths (z_{kh}):

$$var[\varepsilon_k] = \xi z_k \qquad k \in K_{odm} \tag{4.3.22a}$$

$$cov[\varepsilon_k, \varepsilon_h] = \xi z_{kh} \qquad h,k \in K_{odm} \tag{4.3.22b}$$

Usually, variables z_k differ from the actual path cost g_k (e.g. length or uncontested cost). These specifications satisfy the random utility model's property of additivity described in section 3.3.5 and are useful in the analysis of the theoretical properties of equilibrium assignment models to be dealt with in Chapter 5.

Note that the specification (4.3.22) of the variance-covariance matrix of random residuals depends on a single calibration parameter ξ and can be derived by applying the Factor-Analytic Probit model described in section 3.3.6 to the path choice context. As a matter of fact, assuming that a perceived disutility u_l is associated to each link l, with:

$$u_l = \mathrm{E}[u_l] + \eta_l = - c_l + \eta_l$$

where the link random residuals, η_l (l=1, 2, ..., L), are independent normal variables $\eta_l \sim N(0, \sigma_l)$ with:

$$Var[\eta_l] = \sigma_l = \xi r_l$$
$$Cov[\eta_l, \eta_j] = 0$$
$$\eta \sim \text{MVN}(\mathbf{0}, \Sigma_\eta) \qquad \Sigma_\eta = \xi\, \boldsymbol{DIAG}(\boldsymbol{r})$$

where r_l is the link related performance variable corresponding to path attribute z and $\boldsymbol{DIAG}(\boldsymbol{r})$ is the $(n_L \times n_L)$ diagonal matrix containing link variables, r_l. Assuming that the path utility is the sum of its link utilities, it follows that:

$$U_k = \sum_l \delta_{lk} u_l = E[U_k] + \varepsilon_k$$
$$E[U_k] = \sum_l \delta_{lk} E[u_l] = -\sum_l \delta_{lk} c_l = -g_k$$
$$\varepsilon_k = U_k - E[U_k] = \sum_l \delta_{lk}(u_l + c_l) = \sum_l \delta_{lk} \eta_l$$
$$Var[\varepsilon_k] = \sum_l \delta_{lk} \cdot var[\eta_l] = \sum_l \delta_{lk} \cdot \xi r_l = \xi z_k$$
$$Cov[\varepsilon_k, \varepsilon_h] = E[\varepsilon_k, \varepsilon_h] = E\left[\sum_l \delta_{lk}\eta_l \cdot \sum_l \delta_{lh}\eta_l\right] = E\left[\sum_{l \in hk} \eta_l^2\right] = \sum_{l \in hk} var[\eta_l^2] = \xi z_{kh}$$

i.e. the relationships 4.3.22. Since the sum of normal variables is still a Normal variable it results:

$$\varepsilon \sim \text{MVN}(\mathbf{0}, \Sigma)$$

where Σ is the variance covariance matrix with elements given by 4.3.22.

In other words, specification 4.3.22 of the Probit model can be obtained by applying the Factor Analytic Probit to the path choice context with:

$$\varepsilon = \Delta^T \eta = \Delta^T \Sigma_\eta^{1/2} \zeta = \Delta^T [\xi \cdot \boldsymbol{DIAG}(\boldsymbol{r})]^{1/2} \zeta = \boldsymbol{F}\, \zeta$$

where:

- ε is the $(n_p \times 1)$ vector of multivariate normal distributed path random residuals, $\varepsilon \sim \text{MVN}(\mathbf{0}, \Sigma)$;
- Δ is the $(n_l \times n_p)$ link-path incidence matrix;
- η is the $(n_l \times 1)$ vector of independent normal distributed link random residuals, $\eta \sim \text{MVN}(\mathbf{0}, \Sigma_\eta)$;
- ζ is the $(n_l \times 1)$ vector of i.i.d. standard normal random variables, $\zeta \sim \text{MVN}(\mathbf{0}, \boldsymbol{I})$;
- $\boldsymbol{F}$ $\Delta^T \Sigma_\eta^{1/2} = \Delta^T [\xi\, \boldsymbol{DIAG}(\boldsymbol{r})]^{1/2}$ is the $(n_p \times n_l)$ matrix that maps the random vector ζ into path choice random residuals ε;
- n_p is the total number of paths;
- n_l is the total number of links, usually $n_l << n_p$.

It is, in fact, immediate to verify that matrix $\boldsymbol{F}$ specified above, introduced in the (3.3.64) and (3.3.65), gives the (4.3.22a) and (4.3.22b) respectively.

This representation of the Probit path choice model will be used also in section 7.3.1.b for the specification of an algorithm for network assignment to uncongested networks.

The capacity of the Probit model to handle path overlapping, or perceived cost correlation, makes it particularly suitable for applications with exhaustive path generation (implicit enumeration). Furthermore, the difficulty of explicitly calculating the Probit choice probabilities can be overcome with algorithms based on the Monte Carlo simulation described in section 3.3.6. These algorithms will be discussed in Chapter 7.

Recently, a modification to the Logit path choice model has been proposed, called *C-Logit*, which overcomes the problems deriving from Logit IIA property while at the same time retaining an analytical formulation. The C-Logit path choice model has the following specification:

$$p[k/oshdm] = \frac{exp[(-g_k - CF_k)/\theta]}{\sum_{h \in K_{odm}} exp[(-g_h - CF_h)/\theta]} \tag{4.3.23}$$

where the term CF_k, known as the *commonality factor*, reduces the systematic utility of a path proportionally to its level of overlapping with other alternative paths. The commonality factor can be specified in various ways, for example as:

$$CF_k = \beta_o ln\left(1 + \sum_{h \neq k} \frac{z_{hk}}{(z_h z_k)^{1/2}}\right) \tag{4.3.24a}$$

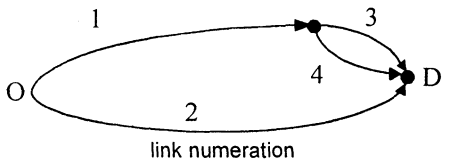
where the attributes z_h, z_k, and z_{hk} are analogous to those described for the Probit model. Expression (4.2.24) shows immediately that the attribute CF_k is inversely proportional to the level of a path's independence, and it is equal to zero if all the links of paths k do not belong to any other path. In this case, it follows that:

$$z_{hk} = 0 \quad \forall h \neq k \rightarrow CF_k = \beta_o \, ln(1) = 0$$

Conversely the attribute CF_k is larger if more paths share the "longer" links of path k. The C-Logit model (4.3.23), for given path costs, reduces the probability of choosing heavily overlapping paths and increases the probability of choosing non-overlapping paths. Furthermore, if the coefficient β_o is equal to 1, C-Logit choice probabilities in the limit case of N coincident paths tend to $1/N$ of those calculated with a Multinomial Logit model applied considering the coincident paths as a single path. These results are illustrated in Fig. 4.3.9, which reports Logit, C-Logit and Probit choice probabilities for a network similar to that in Fig. 4.3.8. As can be seen, C-Logit and Probit probabilities are very similar and lower than those obtained by

the Logit model for heavily overlapping paths. Some calibrations of extra-urban truck path choice models confirm the significance of the attribute CF_k and give β_o values of the coefficient close to one (see Fig. 4.3.7).

	Paths			Link costs
Link	A	B	C	
1	0	1	1	14
2	1	0	0	K
3	0	1	0	2
4	0	0	1	2



K=16

Path	Cost	Logit ($\forall\theta$)	CLogit β_0=1	Probit ξ=1
A	16	0.333	0.478	0.450
B	16	0.333	0.261	0.275
C	16	0.333	0.261	0.275

K=17

		Logit			CLogit			Probit		
Path	Cost	cv=0.1	cv=0.3	cv=1.1	cv=0.1	cv=0.3	cv=1.1	cv=0.1	cv=0.3	cv=1.1
A	17	0.091	0.227	0.302	0.156	0.350	0.442	0.162	0.342	0.421
B	16	0.454	0.387	0.349	0.422	0.325	0.279	0.419	0.329	0.289
C	16	0.454	0.387	0.349	0.422	0.325	0.279	0.419	0.329	0.289

Fig. 4.3.9 Comparison between path choice probabilities with Logit, C-Logit and Probit models.

Expression (4.3.24a) allows computation of the commonality factor additively over the links making up the path and thus application of implicit path enumeration algorithms similar to Dial's (see Chapter 7).

Other specifications of CF have been proposed. Among them, a first one is:

$$CF_k = \beta_o \sum_{l \in k} w_{lk} \ ln \ N_l \qquad (4.3.24b)$$

where the summation is extended to all links l belonging to path k, w_{lk} is equal to the weight of link in path k:

$$w_{lk} = \frac{r_l}{z_k}$$

and N_l is the number of paths between the same O-D pair using link l.

Expression (4.3.24b) takes into account the relative weight of shared links with respect to the overall path cost; for example if two paths h and k, share the same common link l:

$$w_{lh} > w_{lk} \rightarrow CF_k > CF_h$$

The attribute *CF* is larger for a path whose shared links are a larger fraction of its total "length".

Another useful expression of the *CF* is the following:

$$CF_k = \beta_o \, ln\left[1 + \sum_{h \neq k}\left(\frac{z_{hk}}{(z_h z_k)^2} \cdot \frac{z_k - z_{hk}}{z_h - z_{hk}}\right)\right] \qquad (4.3.24c)$$

According to expression (4.3.24c), the *CF* of a path depends also on the cost of non shared links. In this way, the ratio CF_A/CF_B between the commonality factors of two paths increases as the overlapping between two paths (the percentage of common cost with respect to the total one) increases, being $z_A > z_B$.

The C-Logit model has a behavioral interpretation as an Implicit Availability Perception (IAP) model simulating simultaneously the perception of paths as alternatives and the choice among the perceived alternatives as discussed in section 3.4. The commonality factor CF_k, in fact, can be interpreted as an attribute of the model, giving the level of membership of path k to the set of perceived paths I_{odm}, $\mu_{I_{odm}}(k)$:

$$\mu_{I_{odm}}(k) \propto exp(-CF_k) \qquad (4.3.25)$$

i.e, it is assumed that the perception of path k as an elementary alternative is larger if its overlap with other paths is smaller, and vice versa. On the other hand, the first order IAP Logit model described in section 3.4 can be formally expressed as:

$$p[k/odm] = \frac{exp[(-g_k + ln\mu_{I_{odm}}(k))/\theta]}{\sum_{h \in K_{odm}} exp[(-g_h + ln\mu_{I_{odm}}(h))/\theta]} \qquad (4.3.26)$$

Substituting expression (4.3.25) into (4.3.26), expression (4.3.23) results.

4.3.4.2. Path choice models for transit systems

As stated in Chapter 2, public transport systems offer services which are both non-continuous in space (i.e., between discrete points such as stations or stops) and non-simultaneous in time (i.e., available only at times corresponding to departures and arrivals). Supply models (transport networks) representing such systems can be built following two main approaches: "line" based and "run" based. This choice depends on the frequency and the regularity of the service and resulting assumptions on users' behavior. In the following reference will be made to path choice models for scheduled services with frequencies high enough to justify a line-based[17] representation as described in section 2.3.2.1 and repeated in Fig. 4.3.10 for reader's convenience. This assumption is consistent with the within-day stationariety assumption underlying this chapter.

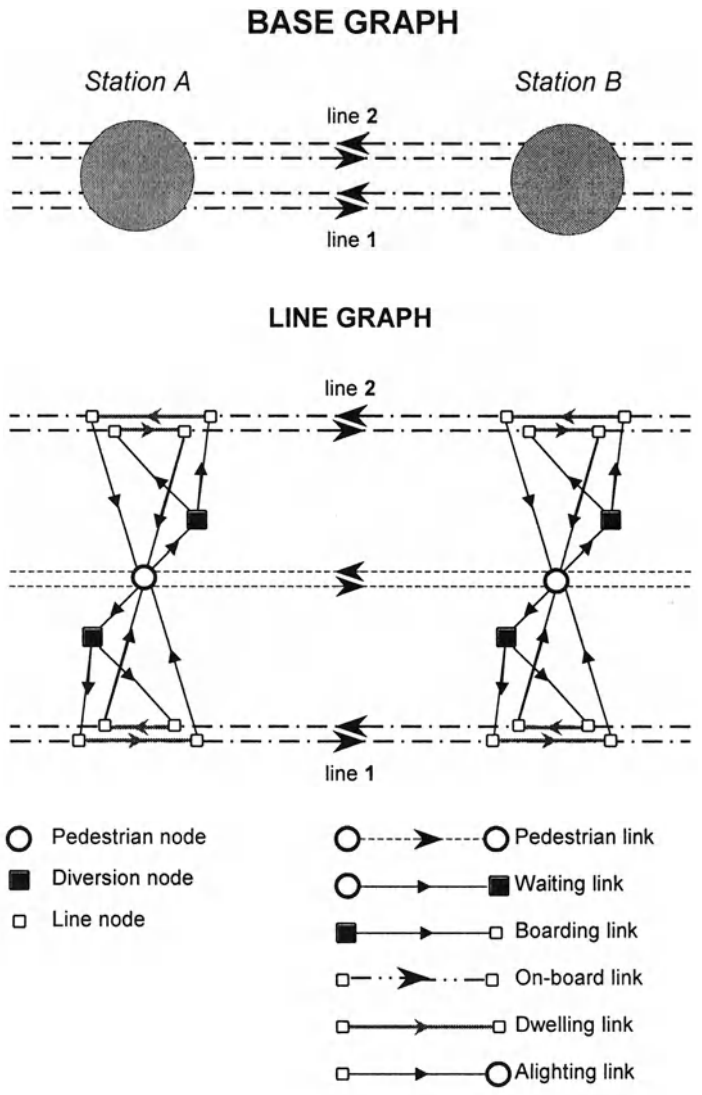


Fig. 4.3.10 Line-based representation of a scheduled transport system.

Also in the case of scheduled service networks, as already said, the complete specification of a path choice model involves three phases: definition of choice alternatives, identification of the set of alternatives and specification of the model simulating choice among alternatives. This in turn implies the definition of the attributes and the systematic utility of the alternatives as well as the functional form of the choice model.

Definition of choice alternatives. In high frequency transport services, it is unrealistic to assume that the user considers as choice alternatives only elementary paths on the graph representing service lines. If this were the case, a user might consider as different and mutually exclusive the paths identified by each of the lines connecting the same pair of stops even when there are several, perfectly equivalent lines. Consider a user traveling in the network represented by the graph of Fig. 4.3.11. If he/she chose the path *b* shown in Fig. 4.3.12 and the line 5 belonging to it, he/she would, on arrival at stop F, refuse to board a vehicle of line 6 arriving at the stop earlier than one of line 5 despite the fact that they are perfectly equivalent. To overcome these potential problems, one should allow for the possibility that the choice alternatives considered by a user before beginning a trip, include several "equivalent" lines, or several paths on the graph representing them. The basic assumption for the definition of choice alternatives is that users of high-frequency transit systems, at the beginning of their trips, do not have complete information. For example, the generic user may be unable to anticipate exactly his/her arrival time at the stops and/or the actual arrival time of the vehicles (trains, buses, etc.) of the different lines calling at each stop. Under this hypothesis it is assumed that the user does not choose a predetermined path but rather a *travel strategy* with the lowest perceived average trip cost. A strategy is defined by a set of pre-defined choices and behavioral rules to follow during the trip, to adapt to random or unknown events. In the example given in Fig. 4.3.11, a strategy could be to go to stop F and board the first vehicle belonging to line 5 or 6; another possible strategy could be to go to stop F and wait only for vehicles of line 5. Two types of behavior are involved in choosing a path under the above assumptions.

En-route choice behavior underlies user choices during the trip. This behavior describes how users respond to unknown or unpredictable events. The type of adaptive choice behavior and the set of alternatives to which it is applied define a strategy.

Pre-trip choice behavior underlies user choices before departure. It includes the comparison of possible alternative strategies and the choice of one of them on the basis of expected characteristics, or attributes. Pre-trip choices are analogous to those assumed for path choice in continuous service networks and, in general, to choices on other dimensions.

The definition of choice alternatives (strategies) therefore requires assumptions about en-route behavior. Usually it is assumed that en-route choices take place at diversion nodes (stops) m and that the en-route behavior rule consist in boarding the first arriving vehicle among those belonging to a given set of lines AL_m, called the set of *attractive lines*[18]. On the other hand, the choice of boarding and alighting stops is made *pre-trip*. To continue the example of Fig. 4.3.11, a strategy cannot include the option of alighting at stop C or at stop D of line 1, because it has been assumed that these are pre-trip choices. This means that there are no events unknown to the user that he/she would adapt to in deciding between either stop. Analogously, a strategy cannot include moving to stop A to take line 1 or to stop B to take line 4.

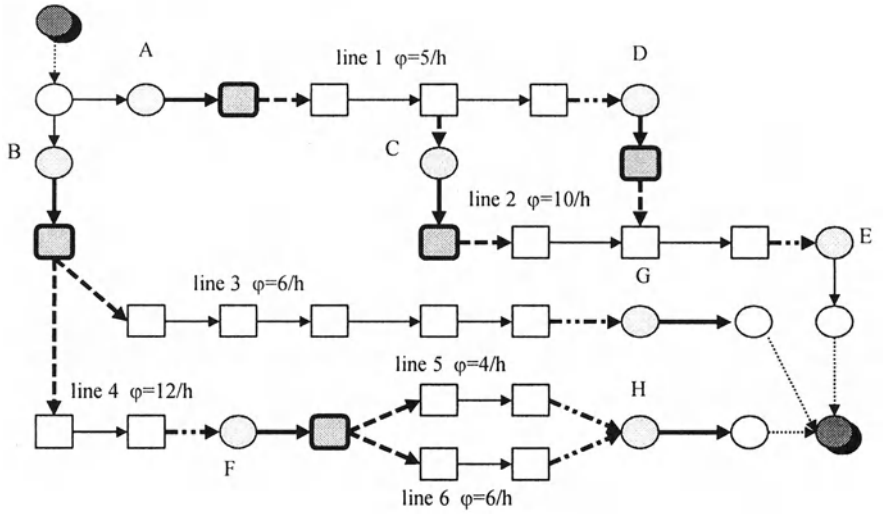


Fig. 4.3.11 Example of a transit line-based network.

If we assume that user and vehicle arrivals at the stops can be modeled as Poisson random processes with homogeneous probability of arrival at any time, the probability of boarding line l belonging to the set of attractive lines at stop m, AL_m, can be expressed as:

$$Pr[l / m, AL_m] = \varphi_l / \sum_{n \in AL_m} \varphi_n \tag{4.3.27}$$

where φ_l represents the frequency (number of arrivals/time units) of line l. Expression (4.3.27) is valid also on the assumption of Poisson arrivals for the user and of deterministic equally-spaced arrivals of the lines belonging to AL_m.

In accordance with these assumptions, a travel strategy, i.e. a pre-trip choice alternative, can be represented by a sub-graph of the line-based graph, known as a *hyperpath.* Elementary paths are theoretically possible strategies, in particular strategies that exclude adaptive choices. Elementary paths are defined *simple hyperpaths*. Strategies that include one or more stops with en-route choices can be represented as the union of simple hyperpaths, such that multiple links can emanate only from diversion nodes[19]. These subgraphs are known as *composed hyperpaths*; Fig. (4.3.12) enumerates all the hyperpaths of the line network in Fig. 4.3.11.

Each diversion node m of hyperpath j will correspond a diversion set AL_{mj} of attractive lines belonging to that hyperpath. To the boarding links $l \equiv (m,n)$ connecting the diversion node m to the nodes n of the lines belonging to AL_{mj}, it is possible to assign a *diversion probability*, η_{lj}. This is the probability expressed by equation (4.3.27) of using the line corresponding to link l of hyperpath j due to the random events underlying en-route choices:

$$\eta_{l,j} = pr[l = (m,n)/\, m, AL_{mj}] = \varphi_l / \sum_{n \in AL_{m,j}} \varphi_n \quad \text{if } l \in AL_{mj} \text{ boarding link} \quad (4.3.28)$$

Typically a diversion probability equal to one is assigned to all non-boarding links belonging to the hyperpath:

$\eta_{lj} = 1$ if $l \in j$, l *non-boarding link*

and a null probability is assigned to the links not belonging to the hyperpath j:

$\eta_{lj} = 0$ if $l \notin j$

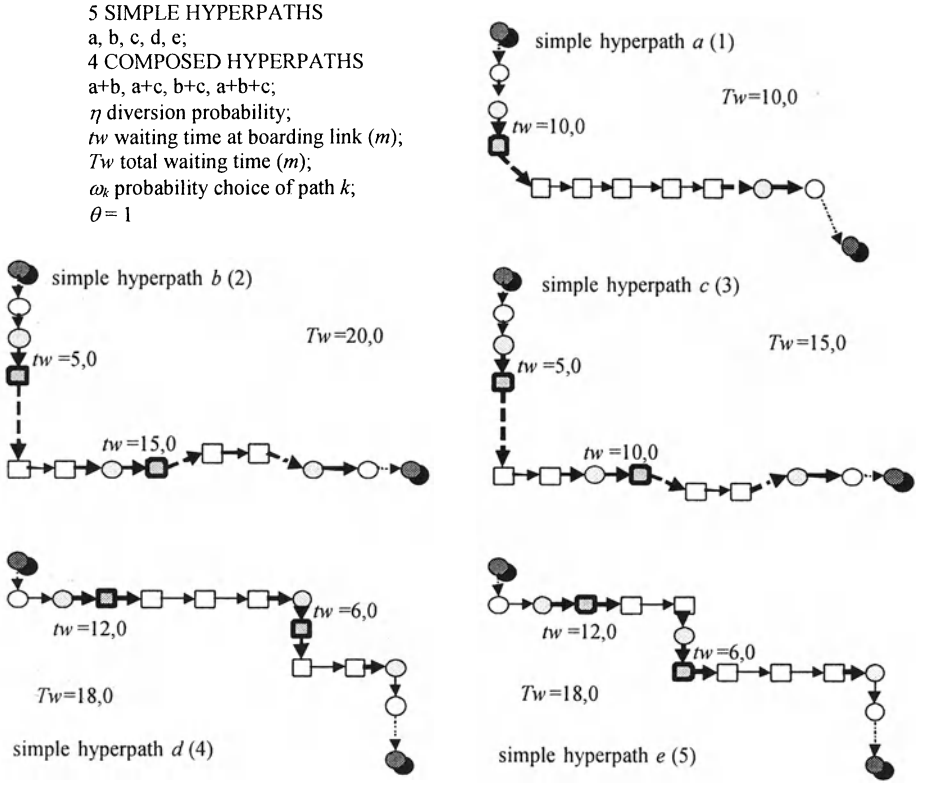


Fig. 4.3.12a Enumeration of simple hyperpaths for the transit network of Fig. 4.3.11.

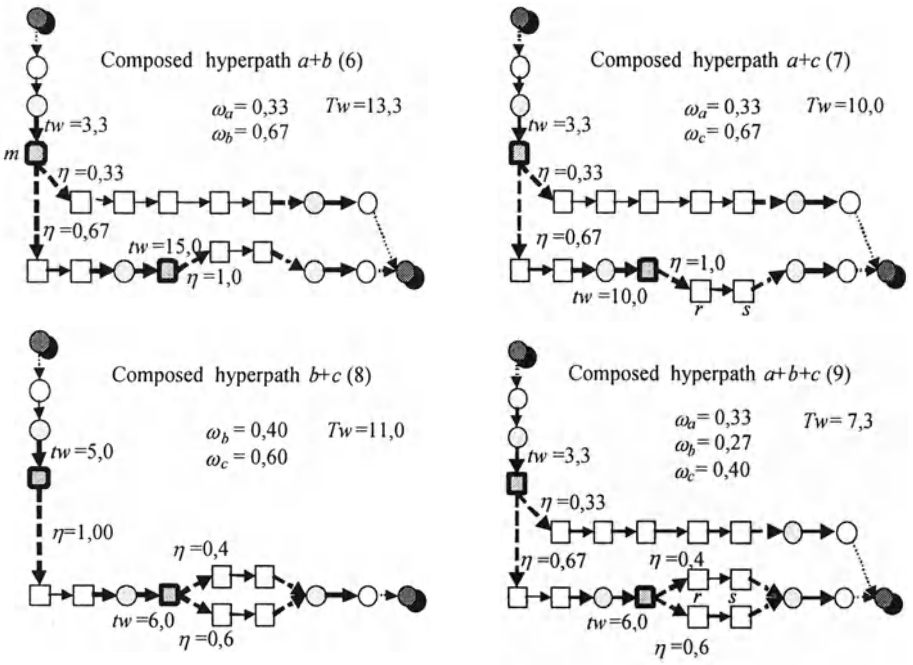


Fig. 4.3.12 b Enumeration of composed hyperpaths for the transit network of Fig. 4.3.11.

For example, the diversion set AL_{m6} corresponding to diversion node m in composed hyperpath 6 in Fig. 4.3.12b consists of the lines 3 and 4: $AL_{m6} = \{3,4\}$ and the diversion probability of boarding link (l) on line 3 can be calculated as:

$$\eta_{l6} = \varphi_3/(\varphi_3+\varphi_4) = 6/18 = 0.33$$

On the basis of the diversion probabilities η_{lj} it is possible to express the probability ω_{kj} of following the path k of the hyperpath j during a given trip. In fact, assuming statistical independence of the random events underlying en-route choices, the probability of following path k within hyperpath j is equal to the product of the diversion probabilities for all links l belonging to path k, i.e.:

$$\omega_{kj} = \Pi_{l \in k}\, \eta_{l,j} \tag{4.3.29}$$

which yields:

$$\omega_{kj} = 0 \qquad k \notin j$$

This probability is obviously equal to one if path k coincides with (simple) hyperpath j. Continuing with the previous example, the probability ω_{a6} of following path a within hyperpath 6 is equal to 0.33; the probability of following the same path

within another hyperpath is different, for example ω_{a1} =1, ω_{a2} =0 and so on. Note that a path may belong to more than one hyperpath. It is also possible to calculate the probabilities λ_{lj} of crossing a link l of hyperpath j as the sum of the probabilities to follow one of the paths k belonging to hyperpath j which includes the link l:

$$\lambda_{lj} = \sum_{k:l \in k} \omega_{kj} = \sum_{k} \delta_{lk}\, \omega_{kj} \tag{4.3.30}$$

which yields:

$$\lambda_{lj} = 0 \qquad \text{if } l \notin j$$

where δ_{lk} is the generic element of the link path incidence matrix. Continuing with the example in Fig. 4.3.12b, the probability of crossing all the links belonging to path b within hyperpath 2 is equal to one; the probability of crossing the link (r,s) is equal to 0.67 in hyperpath 7 and to 0.40 in hyperpath 9. The user choosing a given strategy (or a hyperpath representing it) does not know before starting the trip which path and therefore which lines and links he/she will use since they depend on random events such as the sequence of vehicle arrivals at each stop. On different trips, the same user following the same strategy might use different lines, paths and links with probabilities given by the equations (4.3.28), (4.3.29) and (4.3.20) respectively. Furthermore, on each trip he/she will experience different travel times and, in general, different costs whose mean value can be expressed in function of the probabilities ω_{kj}, as will be shown shortly.

Identification of the choice set. Once choice alternatives (strategies and hyperpaths) have been defined, the issue of the set of such alternatives (choice set) which the user will take into consideration can be considered. As in the case of path choice on road service networks, there are two approaches to the identification of the set of choice alternatives. In the *exhaustive approach*, all strategies (or the hyperpaths that represent them) are feasible. This approach is typically associated with implicit enumeration of the hyperpaths. In the *selective approach*, only those hyperpaths satisfying certain conditions are feasible. For example, hyperpaths including paths with more than one transfer may be excluded from the choice set if there are "direct" paths and hyperpaths. In applications, the most commonly used approach is the exhaustive one, given the calculation complexity of the explicit enumeration of hyperpaths.

Specification of the choice model. Specification of the choice model requires the definition of the attributes and of the functional form of the random utility model. Also, in the case of scheduled service networks, it is assumed that for each hyperpath j belonging to the set $J_{od,m}$ of hyperpaths connecting the pair o,d on the network of the transit mode (or modes) m, the perceived utility of the hyperpath U_j will have a negative systematic utility V_j equal to the mean cost x_j of the hyperpath:

$$U_j = V_j + \varepsilon_j = -x_j + \varepsilon_j \qquad \forall j \in J_{odm} \tag{4.3.31}$$

The average cost of hyperpath x_j can be expressed as the sum of an additive part x_j^{ADD} and a non-additive part x_j^{NA}, which in this case, differently from the path costs on continuous service networks, is always present:

$$x_j = x_j^{ADD} + x_j^{NA} \tag{4.3.32}$$

The additive cost x_j^{ADD} is a linear combination of the attributes (typically on board, boarding, alighting, dwelling and access/egress times) associated with the non-waiting links belonging to the hyperpath:

$$x_j^{ADD} = \beta_b\, Tb_j + \beta_{br}\, Tbr_j + \beta_{al} Tal_j + \beta_d\, Td_j + \beta_a Ta_j \tag{4.3.33}$$

where the β are the respective coefficients.

This cost can be obtained starting from the generalized costs of the single links c_l and the probabilities of crossing the single links (λ_{lj}) or through the additive path costs g_k^{ADD} and the probabilities of following these paths ω_{kj}:

$$x_j^{ADD} = \sum_k \omega_{kj} g_k^{ADD} = \sum_k \omega_{kj} (\sum_{l \in k} c_l) = \sum_l \lambda_{lj}\, c_l \tag{4.3.34}$$

The non-additive cost can be expressed as the sum of the waiting times (costs) Tw_j, as well as any further non-additive costs, i.e. costs which cannot be associated with single links, e.g. fixed fares or transfer costs N_j.

$$x_j^{NA} = \beta_w\, Tw_j + \beta_N N_j$$

where β_w and β_N are the coefficients of reciprocal substitution between the different non-additive cost items.

The average waiting time (cost) Tw_j connected with hyperpath j can be calculated starting from the waiting times tw_{lj} associated with each waiting link l entering diversion node m; as discussed in section 2.3.2.2, this can be expressed as:

$$tw_{lj} = \begin{cases} \theta / \sum_{n \in AL_{m,j}} \varphi_n & \text{if } l \text{ is a diversion link} \\ 0 & \text{otherwise} \end{cases} \tag{4.3.35}$$

where θ is a parameter with values included in the interval [0.5-1] depending on the probability laws of users and vehicle arrivals (see section 2.3.2.2).

The average total waiting cost Tw_j associated with hyperpath j can be expressed as:

$$Tw_j = \sum_{k \in j} \omega_{kj} \left[\sum_{l \in k} tw_{lj} \right] = \sum_l \lambda_{lj} tw_{lj} \tag{4.3.36}$$

From equation (4.3.35) it follows that the waiting time tw_{lj} for the diversion link l depends on the hyperpath and therefore the total waiting time Tw_j cannot be expressed as a linear combination of link attributes independent of the hyperpath; i.e. a non-additive hyperpath attribute.

The choice model among alternative hyperpaths can be expressed formally as the probability q_j that hyperpath j is that of maximum perceived utility:

$$q_j = Pr[-x_j + \varepsilon_j \geq -x_{j'} + \varepsilon_{j'}] \qquad \forall j', j, j' \in J_{od} \tag{4.3.37}$$

Again, in the case of the hyperpath choice model there are two possible approaches. The deterministic choice ($Var[\varepsilon_j]=0$) approach assigns the whole demand to the minimum generalized cost hyperpath(s); alternatively, random utility models, typically Logit and Probit, assign a positive choice probability to all available hyperpaths. When applying the *MNL* model, however, the problems due to the IIA property are even more significant for hyperpaths that include a large number of overlapping lines. Alternatively, it is possible to use a Probit model with a variance-covariance matrix structure similar to that described for paths on road networks.

Unlike path choice models on road networks, in the literature there are no examples of calibration and validation of hyperpath choice models based on observed behavior; this can be explained at least in part by the difficulty of obtaining information on the alternatives (hyperpaths) chosen by users.

Finally, once the hyperpath choice probabilities have been calculated, it is possible to obtain path probabilities:

$$p[k / osdm] = \sum_j \omega_{kj} q_j \tag{4.3.38}$$

4.3.5. A system of demand models

This sub-section describes, as an illustration, the system of extra-urban passenger trip demand models developed and used in the Information System for Transportation Monitoring and Planning in Italy (SIMPT). The system, presented schematically in Fig. 4.3.13, includes models for mobility choices (license holding and number of cars in the family) and partial share trip demand models.

All of the models have a Logit specification and the sequence of frequency/distribution/modal choice models has a three-level Hierarchical Logit structure with inclusive (EMPU) variables taking into account the influence of "lower" choice dimensions on "upper" levels as described in section 4.2. In the following the attributes and their relative interpretations for individual sub-models will be briefly described.

The *driving license holding model* (Fig. 4.3.14) is a Binomial Logit with license holding or not-holding alternatives for each individual. Its systematic utility attributes include the socio-economic characteristics of the individual (age, gender and professional status) and the family (income). The urbanization level of the

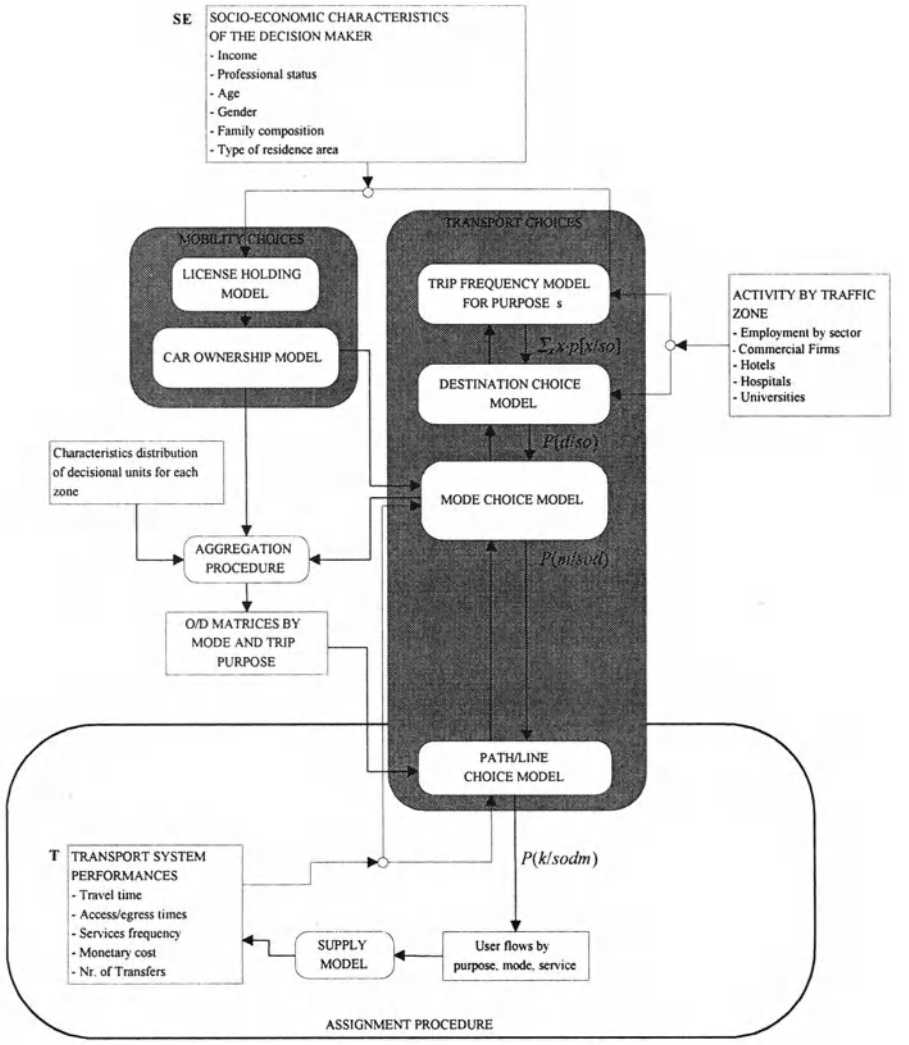


Fig. 4.3.13 Structure of a model system for extra-urban trip demand.

residence zone is also significant. Densely urbanized zones usually have a more efficient public transport system and guarantee better accessibility to various urban functions, reducing the need to use the car. The coefficients indicate that factors such as gender, age, professional status, family income have a significant effect on license holding. Furthermore, it can be observed that the coefficients of socio-economic variables describing gender and age (woman 18-48 and woman > 48) are positive and increasing in the systematic utility of not holding a license. This result can be interpreted as an indicator of the delay with which the female population has gained access to car use, even though this gap is closing for younger generations.

ρ^2=0.437	AGE 18-24 (0/1)	AGE 25-56 (0/1)	EMPLOYED (0/1)	AVERAGE INCOME 40-80 ml	HIGH INCOME >80 ml	DENSE URBAN ZONE (0/1)	WOMAN 18-48 (0/1)	WOMAN >48 (0/1)	ASA
License	0.173	1.146	1.279	0.716	1.229				
t	*2.1*	*16.4*	*19.5*	*10.2*	*5.9*				
No license						0.262	1.197	2.384	-1.022
t						*5.1*	*17.1*	*34.9*	*-16.2*

Fig. 4.3.14 License holding model.

The *car ownership model* (Fig. 4.3.15) simulates the choice of the number of cars owned in the family; the model is a trinomial Logit, with alternatives 0, 1, 2 or more cars. The significant attributes are again socio-economic variables of the family such as income, number of license holders, number of workers and of students. The urbanization level of the residence zone reduces the utility (and the probability) of owning 2 or more cars, confirming the interpretation given for this variable in the license holding model.

ρ^2=0.376	ASA	NR OF WORKERS	NR. OF UNIV. STUD.	FAMILY HEAD (0/1)	AVERAGE INCOME 40-80 ml	HIGH INCOME >80 ml	DENSE URBAN ZONE (0/1)	N°OF LICENSES
0 cars	-1.33	-1.44	-0.99	-0.73				
t	*-13.5*	*-17.2*	*-4.3*	*-7.2*				
1 cars	-0.48							1.06
t	*-24.6*							*27.3*
2 or more cars					1.01	1.53	-0.56	
t					*12.4*	*6.1*	*-6.9*	

Fig. 4.3.15 Car ownership model.

The *trip demand model system* estimates the average number $d^i_{od}[s,h,m,k]$ of extra-provincial round trips undertaken by the generic individual *i* between the zone of residence *o* and the destination *d*, for purpose *s* in the reference period *h*, with mode *m* and path *k*:

$$d^i_{od}[s,h,m,k] = \sum_x xp^i[x/osh](\boldsymbol{SE},\boldsymbol{T}) \cdot p^i[d/osh](\boldsymbol{SE},\boldsymbol{T}) \cdot \\ \cdot p^i[m/oshd](\boldsymbol{SE},\boldsymbol{T}) \cdot p^i[k/oshdm](\boldsymbol{SE},\boldsymbol{T}) \tag{4.3.39}$$

where:

$p^i[x/osh]$ is the probability that individual *i* undertakes *x* extra-provincial trips for purpose *s* in the period *h* obtained with the trip frequency model;

$p^i[d/osh]$ is the probability of choosing destination *d* obtained with the distribution model;

$p^i[m/oshd]$ is the probability of choosing mode *m* obtained with the modal choice model;

$p^i[k/oshdm]$ is the probability of choosing path *k* in mode *m* network obtained with the path choice model.

Five travel purposes are considered: commuting, professional business, study, recreational and tourism, and other purposes.

The *trip frequency model* $p^i[x/osh]$ has a Logit structure with three alternatives: "no trip", "making a single trip", "making more than one trip" in the reference time period h (two winter weeks). The average number of trips undertaken by each individual is therefore obtained as weighted average of the number of trips corresponding to each frequency class (respectively zero, one and the average number estimated by the sample). Weights are given by the probability of choosing each frequency class (see equation 4.3.1). The attributes in the systematic utility functions are the socio-economic characteristics of the family (income level, number of members and cars in the household) and of the traveler (age group, professional status, license holding) and the inclusive utility associated with destination choice $[Y_o^i = ln\Sigma_d\ exp(V^i_{od})]$. Since the model expresses the probability of undertaking journeys external to the province of residence, it includes an "auto-attractivity" variable (e.g. total employment in the province) in the systematic utility of the alternative "no trip". This variable is a proxy for the minor need to carry out activities outside the province for individuals who, everything else being equal, live in areas with more opportunities satisfying their needs. The accessibility variable in the utility of making one or more round-trips has a positive coefficient between zero and one, consistent with the behavioral interpretation of the Hierarchical Logit model. Fig. 4.3.16 shows as an example the attributes and the coefficients calibrated for the "professional business" trip frequency model.

ρ^2=0.7061	TOTAL EMPLOYMENT (x10^6) IN ZONE O	ACCESSIBILITY Y_o^i	AVERAGE INCOME (40-80ml)	HIGH INCOME (>80ml)	MALE (0/1)	MANAGER (0/1)	ASA
0 journeys	0.11						
t	*4.8*						
1 journey		0.14	0.61	1.53	0.96	0.33	-4.80
t		*2.3*	*5.3*	*7.2*	*4.9*	*10.2*	*-13.5*
2 or more journeys		0.14	0.61	1.53	2.34	1.47	-5.592
t		*2.3*			*4.9*	*11.3*	*-14.5*

Fig. 4.3.16 Travel frequency model: for "professional business".

The *distribution model* $p^i[d/osh]$ has a Multinomial Logit specification. Its systematic utility includes the logsum variable Y^i_{od} for mode choice as a (inverse) separation variable between two zones. In order to account for the unknown number of elementary destinations in each zone, "size functions" are used as zone attractiveness attributes, see section 4.3.2. In summary, the utility function of the distribution model can be expressed as:

$$V^i_{od} = \beta_1 Y^i_{od} + \beta_2 ln\left(X^i_{1d} + \sum_{k=2}^{Ks} \beta_k X^i_{kd} \right) + \sum_{k=Ks+1}^{K} \beta_k X^i_{kd}$$

$$with\ Y^i_{od} = ln \sum_m exp(V^i_{odm})$$

where the third term includes all the attributes common to the elementary destinations included in *d*, e.g. the dummy variable "same region" introduced to represent the greater attractiveness, other attributes being equal, of the zones belonging to the same region. The variables included in the "size functions" that depend on trip purpose are service and commerce employment, the number of tourist facilities, and the like.

In the example reported in Fig. 4.3.17 for professional business trips, service employment is used in the "size function" as an indicator of the number of elementary destinations included in each zone. Also in this case, the coefficient of the logsum variable Y^{i}_{od} lies on the [0,1] interval.

ρ^2=0.3129	Y_{od}	SERVICE EMPLOYMENT X_{1d} (x10^3)	SIZE	SAME REGION (0/1)
	0.334	1.000	0.913	1.787
t	*61.3*	-	*13.8*	*42.3*

Fig. 4.3.17 Destination choice model: for professional business.

The *mode choice model* $p^{i}[m/oshd]$ is a Multinomial Logit with six mode or service alternatives: car, bus, air, slow train (inter-regional, express), fast train (inter-city), night train. For each mode, the generic attributes considered are total travel time and monetary cost. In particular, there are two different coefficients for monetary cost, one for low-income users and the other for average-to-high income users. This accounts for different willingness to pay and value of time for users with different incomes, as described in section 4.3.3. The value of time (VOT) perceived by low-income and medium-to-high income users was found to be statistically significantly different. In the example reported in Fig. 4.3.18 for "professional business", the VOT is approximately 5.5 Euro per hour for low-income travelers and 12.5 Euro per hour for medium-to-high income travelers. For "recreational and tourism" and "other purposes", the VOT difference is less dramatic: for medium-to-high income individuals the value of time is approximately 50% higher than for low-income travelers. Other level-of-service attributes are included in the model, such as the number of transfers and the average distance between two runs for scheduled modes/services. For the latter a dummy variable is included, equal to one if the destination zone is not a medium or large city. The negative coefficient of this variable can be interpreted as an (aggregated) measure of the difficulty of reaching the final destination from the service terminal (station, etc.) in the case of low density zones due to less attractive local public transport services. Finally, the specification of the model includes car availability (number of cars divided by the number of licensed drivers in the family) as a socio-economic variable linked to the availability of that alternative.

The *path choice model* on the road network $p^{i}[k/oshdm]$ is also a Multinomial Logit model; the choice alternatives are obtained through an explicit path enumeration technique eliminating heavily overlapping paths. The variables used are

exclusively level-of-service ones. Path choice on scheduled service networks (slow train, fast train, bus and air) is simulated with a Logit model among hyperpaths explicitly enumerated on the lines-based network in accordance with heuristic feasibility rules. Path choice models are applied to origin-destination matrices by mode and trip purpose obtained with the aggregation technique described below.

ρ^2=0.758	TIME [h]	MON. COST LOW INC. [€]	MON. COST MED-HIGH [€]	CAR AVAIL.	NON URBAN DESTIN. (0/1)	NR. OF TRANSF.	TIME HEADWAY [h]	ASA				
								TRAIN			AIR	BUS
								IR	IC	NOTT.		
Car	-1.23	-0.22	-0.098	3.81								
Inter-regional	-1.23	-0.22	-0.098		-3.72	-0.97	-0.60	0.95				
intercity	-1.23	-0.22	-0.098		-3.72	-0.97	-0.60		-0.54			
Night	-1.23	-0.22	-0.098		-3.72	-0.97	-0.60			9.96		
Air	-1.23	-0.22	-0.098		-3.72	-0.97	-0.60				-1.62	
Bus	-1.23	-0.22	-0.098		-3.72	-0.97	-0.60					-2.31
t	*-26.2*	*-5.4*	*-15.7*	*30.3*	*-18.0*	*-5.4*	*-24.0*	*-0.6*	*-4.4*	*3.6*	*-12.7*	*-14.4*

Fig. 4.3.18 Mode and service choice model: for professional business.

The *aggregation procedure* estimates aggregate origin-destination demand flow starting from individual average trips. Since the models described adopt several socio-economic variables at the individual and household level, it would be not feasible to identify user classes characterized by equal values of these attributes. The aggregation procedure adopted is based on the sample enumeration technique described in section 3.7 with the identification of a "prototypical sample" of individuals and families and the calculation of zonal expansion factors calculated so as to satisfy zonal values of aggregate target variables.

4.4. Trip-chaining demand models*

In section 4.1, it was stated that traditional travel demand models simulate the trips making up a journey (sequence of trips starting and ending at home), assuming that the decisions (choices) for each trip are independent of those for other possible trips belonging to the same journey. It has also been said that these assumptions are reasonable when the journey is a "round trip" with a single destination and two symmetric trips.

In recent years, however, there has been an increasing complexity of the structure of human activities, and therefore of travel, especially in urban areas. This has implied an increased number of journeys connecting several activities in different locations, i.e. journeys consisting of sequences of trips influencing each other (Fig. 4.4.1). Consider, for example, the use of car, which cannot be chosen for subsequent trips in the sequence if not used for the first. For these reasons, the literature has proposed several demand models simulating the sequence, or the chain, of trips making up each journey. In particular, some of the models proposed simulate the carrying out of activities (i.e. the purposes of the journey) and the related journey.

The mathematical models proposed to simulate trip chains, or activity chains, do not have a "standard" structure like the case for trip demand models. This is due

both to the recent interest in these models with fewer examples, and to the greater complexity of the phenomenon to be represented.

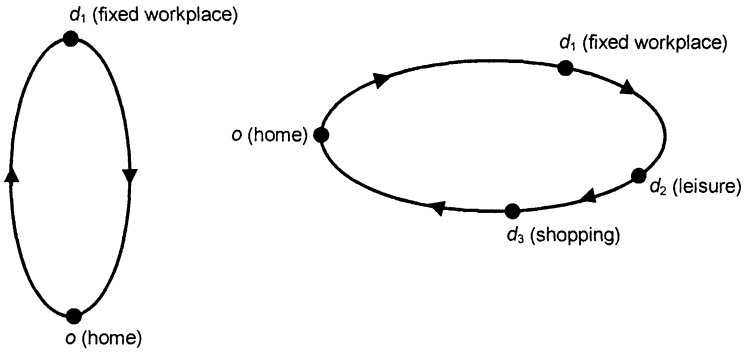


Fig. 4.4.1 Examples of "round trip" and "chain" journeys.

However, the modeling structure most commonly used and closest to that described in the previous sections for single trips is based on the concept of a *primary activity* (*destination*) for a particular journey. In other words it is assumed that each journey can be associated with a primary activity (or purpose), and that this activity is conducted in a particular place, known as the primary destination. Several experimental studies indicate that the activity perceived by the user as being primary for a particular journey can be identified by applying different criteria such as:

- hierarchical level of purpose (in decreasing order, fixed workplace and study, services and professional business, other purposes);
- duration of the activity (the primary activity is that which, within the same hierarchical level, takes most time);
- distance from zone of residence (the primary activity, given the same hierarchical level and duration, is that carried out in the place furthest from the residence).

Adopting this definition, it is possible to specify a system of demand models for trip sequences (journeys) with a partial share structure analogous to the "standard" four-level model described in section 4.2. In order not to excessively complicate the notation, it is assumed that the journeys can have at most two destinations (see Fig. 4.4.2). One of the possible partial share structures for trip chaining is the following:

$$d^i{}_{od_1d_2o}\,[s_1\ h_1\ m_1\ s_2\ h_2\ m_2\ h_3\ m_3] = n^i[o]\ p^i[x{=}1/o\ s_1\ h_1](\boldsymbol{SE}, \boldsymbol{T})\cdot$$
$$p^i[d_1/o\ s_1\ h_1](\boldsymbol{SE}, \boldsymbol{T})\cdot p^i[s_2\,h_2/o\ s\ h_1\ d_1](\boldsymbol{SE}, \boldsymbol{T})\cdot p^i[d_2/o\ s_1\ h_1\ d_1\ s_2\ h_2](\boldsymbol{SE}, \boldsymbol{T})\cdot \quad (4.4.1)$$
$$p^i[h_3/o\ s_1\ h_1\ d_1\ s_2\ h_2\ d_2](\boldsymbol{SE}, \boldsymbol{T})\cdot p^i[m_1\ m_2\ m_3/o\ s_1\ h_1\ d_1\ s_2\ h_2\ d_2\ h_3](\boldsymbol{SE}, \boldsymbol{T})$$

where:

$d^i_{od_1d_2o}$ [s_1 m_1 h_1 s_2 m_2 h_2 m_3 h_3] is the average number of journeys with origin in zone o undertaken by users of category i and composed of trips for primary activity s_1 carried out in zone d_1 in the time period h_1 and secondary activity s_2 carried out in zone d_2 in the time period h_2 and return home in the time period h_3; trips undertaken with modes m_1, m_2 and m_3 respectively. Round trip demand is a special case in which s_2 is return home, d_2 coincides with the origin and m_3 and h_3 are not meaningful;

$p^i[x=1/o\ s_1\ h_1](\boldsymbol{SE}, \boldsymbol{T})$ is the frequency model expressing the probability that an individual of category i resident in zone o undertakes a journey[20] for primary purpose s_1 in the time period h_1;

$p^i[d_1/o\ s_1\ h_1](\boldsymbol{SE}, \boldsymbol{T})$ is the primary destination choice model; it gives the probability that the journey for primary purpose s_1 undertaken by individuals of category i in the time period h_1 has its primary destination in zone d_1;

$p^i[s_2\ h_2/o\ s_1\ h_1\ d_1](\boldsymbol{SE}, \boldsymbol{T})$ is the journey type model; it gives the probability of undertaking a trip for secondary purpose (carrying or not a secondary activity) s_2 in time period h_2 for a user of category i who has decided to undertake a primary journey in d_1 in the time period h_1. Note that the time period h_2 can precede or succeed h_1, i.e., the secondary destination can be reached before or after the primary one, as described in Fig. 4.4.2. Furthermore, if a trip is not undertaken for a secondary purpose, the journey is of the round-trip type and s_2 is the purpose "return home";

$p^i[d_2/o\ s_1\ h_1\ d_1\ s_2\ h_2](\boldsymbol{SE}, \boldsymbol{T})$ is the secondary destination choice model expressing the probability of choosing zone d_2 to carry out activity s_2, if this is not the return home, in the time period h_2 for a user who is undertaking a journey for primary purpose (activity) s_1 in zone d_1 in the time period h_1. This model is obviously meaningless if the journey is a round-trip;

$p^i[h_3/o\ s_1\ h_1\ d_1\ s_2\ h_2\ d_2](\boldsymbol{SE}, \boldsymbol{T})$ is the return home time period distribution model; it gives the probability of returning home in time period h_3 conditional on all the elements that define the chain ($o\ s_1\ h_1\ d_1\ s_2\ d_2\ h_2$) or round-trip ($o\ s_1\ d_1$) journey;

$p^i[m_1\ m_2\ m_3/o\ s_1\ h_1\ d_1\ s_2\ h_2\ d_2\ h_3](\boldsymbol{SE}, \boldsymbol{T})$ is the "mode sequence choice model" for the entire sequence of trips conditional on the elements defining it. Note that mode choice is simulated simultaneously to take into account consistency constraints between successive trips. Some modes (in particular individual modes) are available for successive trips only if they have been used in the first trip.

Path choice models are equivalent to those described in section 4.3.4. It is assumed, in fact, that the probability of choosing a certain path depends exclusively on the origin-destination pair, the mode, and the time period of each single trip without interaction with other trips within the same journey. For this reason, they have not been reported to simplify the analytical formulation and the graphic representation of the models system. Furthermore, ***SE*** and ***T*** denote, as usual, the vectors of socio-economic and level-of-service attributes included in the models.

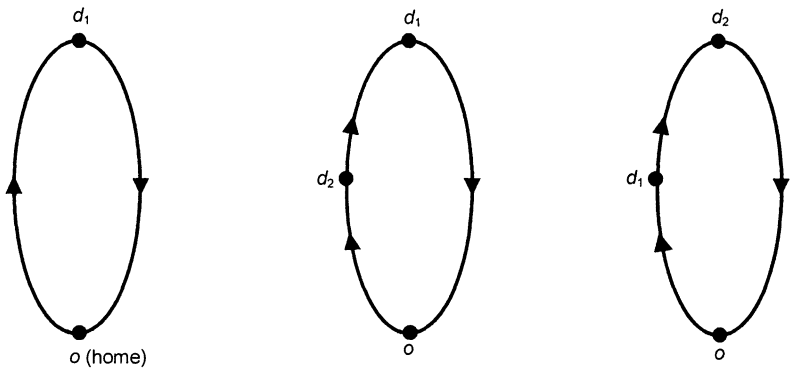


Fig. 4.4.2 Types of journey simulated by the model (4.4.1).

Fig. 4.4.3 is the graphic representation of the structure of the models system described. It can be observed that, similar to the trip demand models system, some choice dimensions are conditional on others, for example the journey type by the primary destination, the secondary destination by the journey type and the primary destination. Upper choice dimensions take into account the lower ones through inclusive or EMPU variables represented by the dotted arrows in Fig. 4.4.3. In the figure, some models in expression (4.4.1) have been further factorialized in the product of two models. In particular, the trip frequency models (primary, secondary and return home) in a certain time period have been decomposed into the product of the probability of undertaking the trip and the probability of choosing a certain time period. The probability of undertaking a return home trip is assumed to be equal to one and is therefore not modeled.

In the context of the partial share structure, different specifications of the whole sequence as well as of individual models can be adopted. Below, a simplified models system for the simulation of trip-chaining travel demand in urban areas is given as an example.

The total model is a Hierarchical Logit, with inclusive logsum variables linking the different choice dimensions; exceptions are the percentages of trips (activities) distribution in the time periods h_1, h_2 and h_3, which are assumed to be constants. The system considers four possible primary purposes: workplace, study, other purposes "constrained" by destination (professional business, personal services, medical treatment, etc.) and other purposes "not constrained" by destination (shopping, recreational, other purposes).

The main models for primary purpose "other non-constrained" are given below. The mode choice model is not included since it is analogous to those described in previous sections; the only significant difference is that the choice alternatives are not single modes/services but their "feasible" combinations, depending on the journey's structure. For example, for round-trip journeys it is assumed that the outward mode coincides with the return mode. For chain journeys it is assumed that if the user employs the car or motorcycle for the first trip, he/she must use it for the

next two, while all combinations of "walking" and public transport modes are possible.

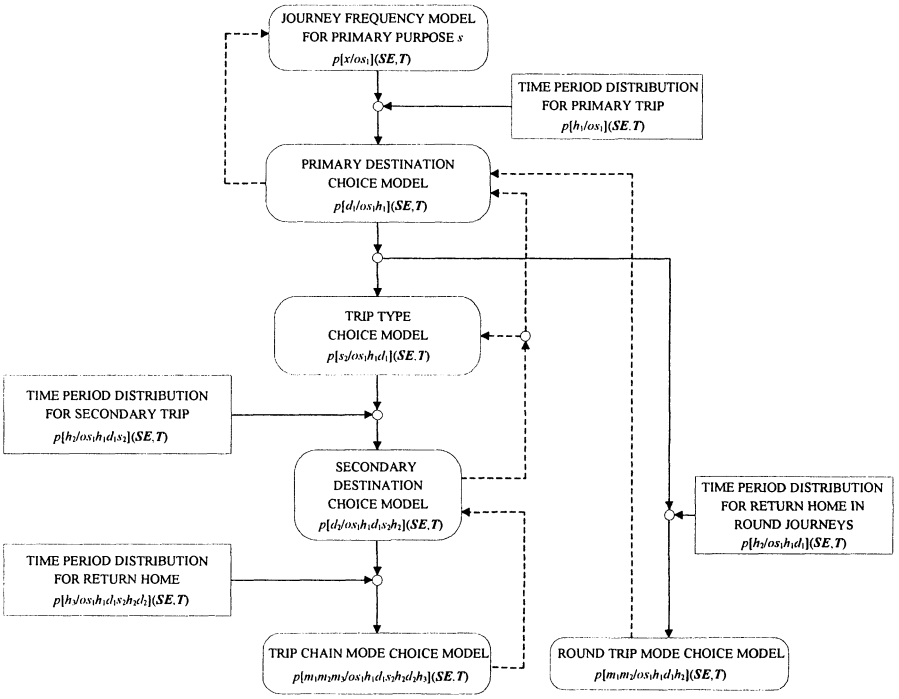


Fig. 4.4.3 Structure of a trip chaining models system.

Journey frequency model $p^i[x/o\ s_1\ h_1](\boldsymbol{SE}, \boldsymbol{T})$. The journey frequency model is a Binomial Logit with systematic utilities of the two alternatives (undertake a journey for the primary purpose or not) given by:

$$V^i_{journey\ s_1} = \beta_1 Y^i_{os_1} + \beta_2 EMP + \beta_3 HSWF + \beta_4 STU + \beta_5 OTHER \qquad (4.4.2)$$

$$V^i_{Nojourney\ s_1} = \beta_6 Nojourney$$

where:

$Y^i_{os_1}$ $= ln\Sigma_{d_1} exp(V^i_{os_1 d_1})$ is the logsum variable corresponding to primary destination choice for the purpose s_1; it represents the accessibility of the residence zone with respect to all the possible destinations where the primary activity can be conducted;

EMP is a dummy variable, equal to one if the individual is of occupational status "employed", zero otherwise;

HSWF is a dummy variable, equal to one if the individual is of occupational status "housewife", zero otherwise;

STU is a dummy variable, equal to one if the individual is an upper school or university student, zero otherwise;

OTHER is a dummy variable, equal to one if the individual is "unemployed" or retired;

NOJOURNEY is the alternative specific attribute (ASA) not to undertake a journey for the primary purpose s_1.

Fig. 4.4.4 reports the parameters calibrated for the described model (4.4.2) for an average weekday. Accessibility of the residence zone increases the probability of undertaking the journey and the logsum inclusive variable has a coefficient on the interval (0,1). The occupational status (category) of the individual considerably influences the probability of undertaking journeys for non-constrained other purposes; employed individuals in particular show, everything else being equal, a lesser utility compared with other categories, probably because of their reduced time budget.

	Y_{os_1}	*EMP*	*HSWF*	*STU*	*OTHER*	*NOJOURNEY*
	0.1904	-0.5879	0.06948	0.5017	0.3607	0.2795
t	*14.6*	*-26.2*	*3.10*	*12.7*	*18.6*	*8.30*

Fig. 4.4.4 Parameters of the journey frequency model for non-constrained other purposes.

Primary destination choice model $p^i[d_1/o\ s_1\ h_1](\boldsymbol{SE},\boldsymbol{T})$. The primary destination choice model is a Multinomial Logit with a systematic utility function of the type:

$$V^i_{od_1s_1h_1} = \beta_1 Y^i_{od_1h_1} + \beta_2 SZ_{d_1/o} + \beta_3 ln(EMPret_{d_1} + \beta_4 EMPserv_{d_1}) \qquad (4.4.3)$$

where:

$Y^i_{od_1h_1}$ $= ln\Sigma_m exp(V^i_{od_1mh_1})$ is the logsum variable corresponding to mode choice and accounting for the (dis)utility to move from o to d_1 using the available transport modes with reference to the user category i and to the departure interval h_1;

$SZ_{d_1/o}$ is a dummy variable equal to one if the zone d_1 coincides with that of residence o, zero otherwise;

$EMPret_{d_1}$, $EMPserv_{d_1}$ are the total employment in retail and service sectors respectively; these variables represent the attractiveness of each primary destination. Since the number of actual elementary destinations in each zone is unknown, this is approximated by means of a "size function" as described in section 4.3.2.

The coefficients reported in Fig. 4.4.5 indicate an increase in the zone's systematic utility as its attractiveness grows. Furthermore, the systematic utility increases as the logsum associated with modal choice increases or decreases the perceived mean cost. Also, the residence zone has an extra-utility, probably due also to the approximations in computing intra-zonal level-of-service attributes.

	$Y_{od_1h_1}$	SZ_{d_1o}	Size	$EMPret_{d_1}(10^3)$	$EMPserv_{d_1}(10^3)$
	1.428	1.003	0.7725	1.000	0.065
t	*19.1*	*9.70*	*19.4*	---	*2.73*

Fig. 4.4.5 Parameters of the primary destination choice model for other unconstrained purposes.

Journey-type choice model $p^l[s_2/o\ s_1\ h_1\ d_1\ h_2](\boldsymbol{SE},\boldsymbol{T})$. This model simulates the choice between two alternatives: undertaking a further trip for a secondary purpose (trip-chain journey) or return home (round-trip journey). The model is therefore a Binary Logit with the following systematic utility functions:

$$\begin{aligned} V_{chain} &= \beta_1\, ML + \beta_2\, EMP + \beta_3\, STU + \beta_4\, OTHER + \beta_5\, MRNG + \\ &\quad + \beta_6\, AFTN + \beta_7\, EVNG \\ V_{round} &= \beta_8\, ROUND + \beta_9\ DACC_{od_1} \end{aligned} \qquad (4.4.4)$$

where:

ML is a dummy variable, equal to one if the individual is male, zero otherwise;

EMP is a dummy variable, equal to one if the person is employed, zero otherwise;

STU is a dummy variable, equal to one if the person is an upper secondary school or university student, zero otherwise;

OTHER is a dummy variable, equal to one if the person is a housewife, retired, unemployed, zero otherwise;

MRNG is a dummy variable, equal to one if the trip starts before 12.00 (h_1<12), zero otherwise;

AFTN is a dummy variable, equal to one if the trip starts between 12.00 and 16.00 (12<h_1<16), zero otherwise;

EVNG is a dummy variable, equal to one if the trip starts between 16.00 and 20.00 (16<h_1<20), zero otherwise;

ROUND is the Alternative Specific Attribute for the alternative "round trip";

$DACC_{od_1}$ is the accessibility differential of the residence zone *o* and primary destination zone d_1; accessibilities are calculated as logsum variables with respect to destination choice for the considered purpose.

The coefficients obtained from the calibrations are reported in Fig. 4.4.6. As can be seen, employees and students have, everything else being equal, a greater utility for chained trips, most likely because of their limited time budget. There is a larger systematic utility for, and therefore a larger probability of undertaking chain trips during the morning than during the evening and, even more, than during the afternoon. The role of the accessibility attribute $DACC_{od}$ deserves some further comment. If a residence zone has a larger accessibility with respect to the possible locations satisfying mobility needs for "other unconstrained purposes" than the primary destination, the return home probability increases; on the other hand, if the residence zone has a lower accessibility, the probability of undertaking a chain trip increases. Schematically, it is more likely, everything else being equal, that a person who lives in the suburbs and undertakes a primary trip to the city center, will undertake a trip chain than the opposite case in which the person, once home, can undertake another journey to satisfy his/her further needs.

	ML	EMP	STU	OTHER	MRNG	AFTN	EVNG	ROUND	$DACC_{od_1}$
	1.708	0.4185	1.107	-0.3559	0.5295	-1.311	0.1835	4.4640	0.3934
t	*24.8*	*7.30*	*11.10*	*-4.80*	*8.60*	*-11.9*	*3.10*	*61.9*	*7.8*

Fig. 4.4.6 Parameters of the journey type choice model for "other unconstrained" purposes.

Secondary destination choice model $p^i[d_2/o\ s_1\ h_1\ s_2\ h_2]$**(SE,T)**. The secondary destination choice model is a Multinomial Logit with systematic utility functions similar to those described for the primary destination choice model:

$$V^i_{od_2} = \beta_1 Y^i_{d_1 d_2 o h_2} + \beta_2 ZN_o + \beta_3 ln(EMPret_{d_2} + \beta_4 EMPserv_{d_2}) \qquad (4.4.5)$$

where:

$Y^i_{d_1 d_2 o h_2}$ = logsum inclusive variable of the mode choice model accounting for the (dis)utility of all modes from primary destination d_1 to secondary potential destination d_2 and to residence zone o;

$SZ_{d_2/o}$ = dummy variable equal to one if the zone d_2 coincides with that of residence o, zero otherwise;

$EMPret_{d_2}$, $EMPserv_{d_2}$ = total employment in the retail and service sectors respectively included in the "size function" which expresses the attractiveness of zone d_2 as a potential secondary destination.

The coefficient estimates, reported in Fig. 4.4.7, align with expectations indicating, everything else being equal, a larger utility for secondary destinations with lower generalized transport cost and larger attraction capacity (greater number of elementary destinations).

	$Y_{d_1 d_2 o h_2}$	$SZ_{d_2 / o}$	Size	$EMPret_{d_2}(10^3)$	$EMPser_{d_2}(10^3)$
	0.417	1.865	0.684	1.0000	0.618
t	*2.90*	*5.00*	*3.80*	---	*1.0*

Fig. 4.4.7 Parameters of the secondary destination choice model for "other unconstrained" purposes.

4.5. Applications of demand models

To conclude the analysis of traveler demand models, it is useful to comment on the "nature" of their application and on their fields and modalities of application.

The "true" values of demand flows (present and predicted) are generally unknown to the analyst and as such must be represented as random variables. Demand models provide possibly unbiased estimates of average demand flow values with certain characteristics. In some cases it is also possible to compute variances and covariances of the estimates obtained. For example, with reference to the case of a four-level demand model and a single trip for each purpose s in reference period h the demand flow $d_{od}[s,h,m,k]$ can be modeled as a multinomial r.v. In other words, the demand estimates obtained with a partial share model are the mean (expected) values of random variables which, assuming statistical independence of individual decisions, can be assumed to be distributed with a multinomial law. It is therefore possible to express the variances and covariances of demand flows obtained from the models:

$$
\begin{aligned}
&E[d_{od}\,[shmk]] = n[osh]\; p[xdmk/osh] \\
&Var[d_{od}\,[shmk]] = n[osh]\; p[xdmk/osh]\; [1\text{-}p[xdmk/osh]] \\
&Cov[d_{od}\,[shmk]\; d_{od'}\,[shm'k']] = n[osh]\; p[xdmk/osh]\; p[xd'm'k'/osh]
\end{aligned}
\qquad (4.5.1)
$$

The actual "deviation" of the estimates obtained with the models with respect to the "true" demand flows is certainly larger than that expressed by the variance (4.5.1). In fact, models, however sophisticated, are only simplified representations of the complex phenomena underlying mobility, and therefore the probabilities $p[xdmk/osh]$ are only estimates of real percentages whose deviation (variance) can only be calculated on an empirical basis.

The practical uses of demand models can be divided into three categories: estimation of present demand and its variations, quantitative analysis of the characteristics of mobility, and components of the system of demand-supply interaction, i.e. assignment models. These three application typologies imply several requirements of the models that will be briefly dealt with below.

Estimation of present demand and its variations. This is the "classic" use of demand models. The models can be used as estimators of present demand, i.e. as mathematical structures underlying transport demand which, once specified and calibrated, are applied to present activity and transportation supply systems to estimate unknown demand flows. Alternatively the models can be used to simulate (or "forecast") variations in travel demand induced by changes in the activities

and/or transportation supply systems. For both these applications, different techniques can be used depending on the application context and the models can be integrated with other information available.

Application of the models both to estimate present demand and to simulate its variations requires the results to be aggregated in order to obtain estimates of demand flows between different origin-destination pairs. For this function, the different *aggregation techniques* described in section 3.7 for random utility models can be used, depending on the characteristics of the models specified. Aggregate models refer to aggregation techniques by category, implicitly assumed in expression (4.2.2), while disaggregate models can be used in conjunction with sample enumeration techniques with target variables corresponding to the present situation or predicted for a future scenario. These topics will be dealt with in more detail in Chapter 8.

Tools for quantitative analysis of mobility. Another possible use of demand models is as a *statistical tool for quantitative analysis* of mobility phenomena. In this case the models are seen as relationships allowing the opportunity to evaluate quantitatively the influence of both socio-economic and level-of-service variables on mobility. In this case the emphasis is not on the application of the models to obtain aggregate demand estimates (present or future) but on the specification and the estimation of the coefficients of the model itself.

Some of the models described in this chapter could be used, for example, for the quantitative analysis of the influence of factors such as age, sex, income, occupational status, etc. on the different aspects of mobility that have been examined. For this type of application, the variables used might be very detailed since it is not necessary to know their present values over the whole universe or their predict future values.

Demand models for assignment to transport networks. Demand flows obtained with models are often used as input for assignment models to simulate flows and performance of various elements of the supply system represented by the links of a transportation network. For this type of application, the models are considered to be demand functions. They express origin-destination flows with different modes during the reference period as a function of socio-economic variables $\boldsymbol{SE}$ and of generalized route costs $\boldsymbol{g}$. On the other hand, route choice models are explicitly used for the formulation of assignment models.

For the formulation of assignment models, demand models are represented with a notation slightly different from that used so far. Since the demand-supply interactions simulated with assignment models relate at least to route choice, the route choice model is separated from those on other levels (choice dimensions). In this case, the generic partial share model becomes

$$d^i_{od}[hmk](\boldsymbol{SE},\boldsymbol{T}) = d^i_{od}[hm](\boldsymbol{SE},\boldsymbol{T})\, p^i{}_{od,k}(\boldsymbol{g}^i{}_{od,m})$$

where, as will be seen in more detail in Chapter 5, $\boldsymbol{g}^i{}_{od,m}$ is the vector of generalized route costs corresponding to the *od* pair on the mode *m* network and to user class *i*

and the attributes corresponding to route choice different from those contributing to the generalized transportation cost are implied. As stated in section 4.3.4 on route choice models, path cost coincides with the opposite of systematic utility, $V_k = -g_k$. Generalized route costs $\boldsymbol{g}^i_{od,m}$, homogenize different attributes which are components of the vector $\boldsymbol{T}$. It should also be noted that trip purpose s does not appear explicitly in the previous expression since, in the assignment context, the index i will denote the class of users defined by the pair (category, purpose)[21].

Furthermore, in assignment models the aggregated O-D flow for user class i is denoted by d^i_{odm} if the demand is considered rigid, i.e. not sensible to variations of generalized costs due to network congestion. If the demand is considered elastic on some or all dimensions, the demand function is denoted by $d^i_{odm}(\boldsymbol{s}(\boldsymbol{g}))$. In the case of elastic demand, models simulating variations on other dimensions use the EMPU variable $s_{m/od}$ corresponding to route choice on the mode network m in time period h for users of class i which depends on the costs of the different routes. The EMPU variables corresponding to all O-D pairs can be ordered in the column vector $\boldsymbol{s}_m$. The different notation between demand flow d^i_{odm} and demand functions $d^i_{odm}(\boldsymbol{s}(\boldsymbol{g}))$ does not imply that the latter cannot be obtained with the demand models described in this chapter; it rather underlines the dependence of demand on congestion-related costs in the analysis of interactions between elastic demand and supply (elastic demand assignment models). This notation will be taken up in more detail in Chapter 5.

4.6. Freight transport demand models*

Freight transport demand is closely connected to the production and distribution of goods, i.e. to the economic system under study and to its interactions with external economic systems. Many of the definitions and classifications presented for passenger transport demand can be extended to freight transport demand, although their interpretation is in general very different. A system of freight demand models can be formally expressed as:

$$d_{od}\,[K_1, K_2, \ldots] = d\,(\boldsymbol{SE}, \boldsymbol{T}, \beta) \tag{4.6.1}$$

although, beyond the formal analogy with expression (4.1.1) for passenger demand models, the interpretation of the symbols is significantly different. Demand flows represent movements of quantities of freight (usually expressed in tons); the relevant characteristics K_1, K_2,..., are normally associated with *goods typology* (raw materials, semi-finished products, finished products, etc.), with *economic activity sectors*, with *industrial logistics characteristics* (e.g. shipping frequency and size) as well as with *modes of transport*. The latter are usually defined not only by the physical vehicle (truck, train, ship) but also by their organization (own shipment, by carrier, etc.). The $\boldsymbol{SE}$ variables are those of the economics of production (value of production by sector, number and size of local units, etc.) and consumption (household consumption, imports, etc.); the variables of the transport system, $\boldsymbol{T}$, are still related

to the attributes of the different transport modes and services (times, costs, service reliability, etc.). The vector β denotes the coefficients in the model and will be understood as given in what follows.

These considerations suggest that the mechanisms underlying the formation of freight transport demand and its satisfaction by transport services are considerably more complex and articulated than those corresponding to passengers. In the case of freight, in fact, there is not a single decision-maker (the individual) but rather a complex and articulated set of decision makers responsible for production activities, logistics (storage and shipping), product distribution and marketing.

Schematically, it is possible to group the decision-makers who influence the level and composition of freight transport demand into three categories. *Producers* of goods and services decide how much and how to produce, and where and at what prices to sell; *consumers,* either intermediate (production companies) or final (families, public administration), decide how much and how to consume; and *shippers* (transport companies) decide how to provide transport services.

Some classification factors of demand models proposed for passengers can be extended to freight.

The models can be *disaggregate* or *aggregate* depending on whether variables are measured in disaggregate units such as individual companies or individual shipments, or in aggregate units such as all the companies of a certain category and/or economic sector. Furthermore, freight demand models can be *behavioral* or *descriptive* depending on whether they are based on explicit assumptions regarding the behavior of market agents or on empirical relationships between freight transport demand and causal variables corresponding to the economic and/or transport system.

Freight transport demand models have been studied and applied to a lesser extent than passenger models, mainly because of the complexity of the underlying phenomena that influence freight transport. In any case, there is not a consolidated paradigm but rather only some examples, which depend on the type of application and the data available.

The most recent and sophisticated systems of models for freight demand simulation have resulted from the integration of two classes of models: macro-economic models which simulate the level (quantity) and spatial distribution of goods exchanged between different economic zones (leading to origin-destination matrices), and models which simulate mode and route choice.

Within these common characteristics, two main groups of freight demand models have been proposed, as well as different specifications within each group dependent on the variables explicitly simulated. Models of the first group, known as *Spatial Price Equilibrium* models (*SPE*), simulate the production and consumption of each zone and each economic sector through demand and supply curves as functions of prices; they also assume deterministic demand behavior in that there is commercial exchange of goods between two zones only if the sale price in the origin (production) zone plus the transport cost is equal to the sale price in the destination (consumption) zone. The problem of equilibrium of prices, quantities exchanged, and transport costs can be formulated under certain assumptions as a non-linear

programming problem with linear constraints. *SPE* models can be extended and generalized in several directions, though they have received some criticism basically concerning the lack of realism of the deterministic assumptions. This leads to demand flows that are "polarized" towards few origin-destination pairs and zero towards others (contrary to all the experimental evidence), and the use of zonal demand and supply functions which do not take into account the interdependence of various economic sectors.

The second group of models originates from an explicit representation of the interdependence between the different sectors of economy to simulate the quantity of goods produced and exchanged between different zones (intersectorial models). Within this group can be placed various models that, as will be seen soon, differ from each other with respect to the elements of the economic system that are considered rigid or elastic, and with respect to their implicit or explicit simulation of the price system. Models of this type are usually coupled with random utility mode choice models which can be aggregate or disaggregate according to the attributes in their systematic utility functions. Route choice models, at least in the case of road networks, are completely analogous to those described for passengers.

In the following, the general structure of Multi-regional Input-Output intersectorial models will be described (section 4.6.1). These models are among the most flexible and generalizable formulations for the simulation of freight demand level and spatial distribution. Some models for freight transport mode choice will be described in section 4.6.2. Examples of both models will be given drawing from the integrated system of models used to simulate freight demand in Italy, whose structure is represented in Fig. 4.6.1.

4.6.1. Multiregional Input-Output (MRIO) models

The application of macro-economic models to freight demand simulation is usually conducted in two phases, as illustrated in Fig. 4.6.1. In the first phase, the exchange (or trade) between economic sectors and regions is simulated "in value", i.e. in monetary terms, while in the second phase monetary exchanges are transformed into quantity exchanges (tons). The latter results in the input O-D matrices for mode/service and path choice models.

In the first phase it is possible to use a variable coefficient multi-zonal sectorial inter-dependence model, referred to as a Multi-Regional Input-Output (MRIO) model.

It is assumed that the study area is divided in n_z zones in accordance with the principles for zoning described in Chapter 1. It should be noted that in applications of macro-economic models relatively large zones are used; this is due to the availability of statistical information needed by the model. Typically zones coincide with entire regions, from which the name MRIO derives. The transition to a finer zoning system, necessary for the simulation of mode choice and networks assignment, can be conducted in the second phase, where values are transformed into quantities, e.g. using descriptive demand models.

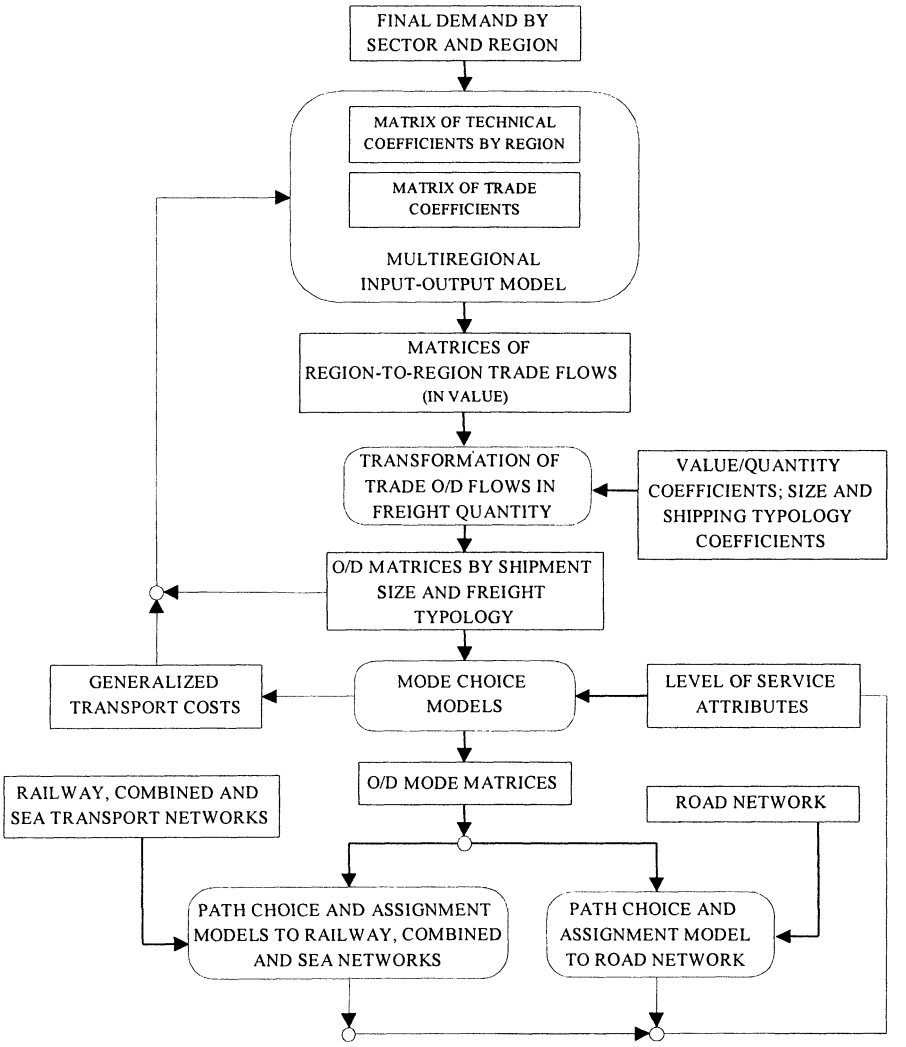


Fig. 4.6.1 Structure of a model system for freight transport demand.

Economic activities of production and consumption are divided into n_s sectors. These can be sectors of the economy producing goods (e.g. agriculture and industrial sectors) or services (e.g. banking and commerce). The activities within each sector are considered to be homogeneous with respect to their economic behavior. A large number of sectors would guarantee an accurate description of relevant economic phenomena and greater plausibility of the assumption of behavioral homogeneity; on the other hand, in applications it is necessary to take into account the aggregation

levels of the available data. Fig. 4.6.2 shows the 17 macro-sectors used to represent the Italian economy for the above-mentioned system of national models.

		SECTORS
GOODS MANUFACTURING SECTORS	1	Agriculture, forestry and fisheries
	2	Energy products
	3	Ferrous and non-ferrous minerals and metals
	4	Non metalliferous minerals and products
	5	Chemical and pharmaceutical products
	6	Metal products and machinery
	7	Means of transport
	8	Foodstuffs, drinks and tobacco
	9	Textile products, clothing, leather goods and footwear
	10	Paper, paper products, printing and publishing, other industrial products
	11	Wood, rubber
SERVICE SECTORS	12	Buildings and civil engineering
	13	Commerce, hotels and public utilities
	14	Transport and communication
	15	Banking and insurance
	16	Other services for-sale
	17	Services not for sale

FINAL DEMAND COMPONENTS
Family consumption
Public consumption
Investments
Stock variations
Export

Fig. 4.6.2 Sectors of the economy and components of final demand for the national model.

As stated, exchanges between sectors and zones (and every other variable homogeneous with them) are expressed by their corresponding economic value measured in monetary units, usually with reference to a given year.

The demand for products (goods and/or services) is assumed to be subdivided into two parts. They are respectively the *intermediate demand*, i.e. production used as inputs to further production in the same or in other sectors within the study area and the *final demand*, i.e. the production used for final consumption internal or external to the study area (export). For example, part of the production of the engineering industry (industrial machinery) may be used to produce goods within the same sector or used in other industrial sectors (e.g. the textile industry). Another part may be used for the production of services (e.g. as office equipment) and yet another may be used for final consumption (e.g., washing-machines for family use or other machines exported outside the study area).

By definition, final demand consists of all uses of sector production which are not re-employed for production within the study area; it usually includes several uses: families and institutions consumption, stock variations and export. Fig. 4.6.2 illustrates the elements of final demand taken into account in the national model.

To give a formal description of the MRIO model, it is necessary to introduce some new variables which are described below. Let:

- Z^{mn}_{ij} be the value of the intermediate demand of sector m produced in zone i and necessary for the production of (consumed in) sector n in zone j;
- K^{mn}_{j} be the value of the intermediate demand of the production in sector m necessary for the production of sector n in zone j, with $K^{mn}_{j} = \Sigma_i Z^{mn}_{ij}$;
- W^{m}_{ij} be the value of sector m produced in zone i necessary to satisfy the final demand in zone j;
- Y^{m}_{j} be the value of the final demand in sector m in zone j. From the definitions given, it results: $Y^{m}_{j} = \Sigma_i W^{m}_{ij}$;
- $\boldsymbol{Y}$ be the vector of final demand of dimensions $(n_z \cdot n_s \times 1)$ obtained by ordering the elements Y^{m}_{j} for each sector and for each region;
- X^{m}_{i} be the value of the total production of sector m in zone i;
- $\boldsymbol{X}$ be the vector of the total production of dimensions $(n_z \cdot n_s \times 1)$;
- J^{m}_{i} be the value of imports of sector m in zone i;
- $\boldsymbol{J}$ be the vector of imports of dimensions $(n_z \cdot n_s \times 1)$.

All variables are expressed in monetary flow units (e.g. million EUROS per year).

Since by definition the total supply (production and import) of sector m in zone i must be used for production in other sectors or consumed (including export), it must be equal to the total demand (intermediate and final) of sector m produced in zone i. The latter is given by the sum of the intermediate demand (in any zone and any sector) and of the final demand (in any zone):

$$X^{m}_{i} + J^{m}_{i} = \Sigma_j \Sigma_n Z^{mn}_{ij} + \Sigma_j W^{m}_{ij} \tag{4.6.2}$$

The following relationship between the values of production and of intermediate demand can be established:

$$K^{mn}_{j} = a^{mn}_{j} X^{n}_{j} \tag{4.6.3}$$

where the *technical coefficients* a^{mn}_{j} represent the value of the product of sector m (input) necessary to produce a unit of value of sector n (output) in zone j. These coefficients depend on the production "technologies" available in zone j; in general, the lower the coefficients a^{mn}_{j}, the more efficient the production in j since a lesser input value is required for an output unit. The elements a^{mn}_{j} corresponding to a given region j can be ordered in a square matrix $\boldsymbol{A}_j$ $(n_s \times n_s)$ known as the *matrix of technical coefficients of region j*. Different regions can have different production technologies and technical coefficient matrices. The matrices $\boldsymbol{A}_j$ can be arranged in a block diagonal matrix $\boldsymbol{A}$ of dimensions $(n_z \cdot n_s \times n_z \cdot n_s)$, in which each block relates to a zone. Fig. 4.6.3 reports an example of some of the variables introduced, corresponding to a 3-region, 2-sector system (market).

Note that the relations (4.6.1)-(4.6.3) do not introduce any modeling hypothesis and can be considered a "re-organization" of the data on present economic exchanges between sectors and regions. Usually not all of the information related to the variables introduced is available, and this data would take on different values in the future thus limiting the value of present information. In order to estimate the parameters of a model and apply it to scenario forecasts, it is therefore necessary to introduce some simplified hypotheses from which different formulations of Sectorial Interdependence models derive. The Input-Output model will be shown to be a special case.

MRIO model with constant coefficients. The first simplified hypothesis, introduced by Cenery-Moses, is to assume that the acquisition percentages of the product of sector m in zone i are independent of the sector n in which this product is employed. In other words, it is assumed that we can express the acquisition needs of zone i products of sector m for production of sector n in zone j (Z^{mn}_{ij}), or for the final consumption (W^{m}_{ij}) as:

$$\begin{aligned} Z^{mn}_{ij} &= t^{m}_{ij} \cdot K^{mn}_{j} \\ W^{m}_{ij} &= t^{m}_{ij} \cdot Y^{m}_{j} \end{aligned} \tag{4.6.4}$$

Vector of sectorial production $\boldsymbol{X}$ (3·2×1)

REGION A	Sector 1	X^{1}_{A}
	Sector 2	X^{2}_{A}
REGION B	Sector 1	X^{1}_{B}
	Sector 2	X^{2}_{B}
REGION C	Sector 1	X^{1}_{C}
	Sector 2	X^{2}_{C}

Matrix of the technical coefficients $\boldsymbol{A}$ (3·2×3·2)

		REGION A		REGION B		REGION C	
		Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2
REGION A	Sector 1	a^{11}_{A}	a^{12}_{A}	0	0	0	0
	Sector 2	a^{21}_{A}	a^{22}_{A}	0	0	0	0
REGION B	Sector 1	0	0	a^{11}_{B}	a^{12}_{B}	0	0
	Sector 2	0	0	a^{21}_{B}	a^{22}_{B}	0	0
REGION C	Sector 1	0	0	0	0	a^{11}_{C}	a^{12}_{C}
	Sector 2	0	0	0	0	a^{21}_{C}	a^{22}_{C}

Fig. 4.6.3a Variables for a 3-region, 2-sector MRIO model.

Matrix of exchange or trade coefficients $\boldsymbol{T}$ (3·2×3·2)

		REGION A		REGION B		REGION C	
		Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2
REGION A	Sector 1	t^1_{AA}	0	t^1_{AB}	0	t^1_{AC}	0
	Sector 2	0	t^2_{AA}	0	t^2_{AB}	0	t^2_{AC}
REGION B	Sector 1	t^1_{BA}	0	t^1_{BB}	0	t^1_{BC}	0
	Sector 2	0	t^2_{BA}	0	t^2_{BB}	0	t^2_{BC}
REGION C	Sector 1	t^1_{CA}	0	t^1_{CB}	0	t^1_{CC}	0
	Sector 2	0	t^2_{CA}	0	t^2_{CB}	0	t^2_{CC}

O/D matrix of value exchanges $\boldsymbol{N}$ (3·2×3·2)

		REGION A		REGION B		REGION C	
		Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2
REGION A	Sector 1	N^{11}_{AA}	N^{12}_{AA}	N^{11}_{AB}	N^{12}_{AB}	N^{11}_{AC}	N^{12}_{AC}
	Sector 2	N^{21}_{AA}	N^{22}_{AA}	N^{21}_{AB}	N^{22}_{AB}	N^{21}_{AC}	N^{22}_{AC}
REGION B	Sector 1	N^{11}_{BA}	N^{12}_{BA}	N^{11}_{BB}	N^{12}_{BB}	N^{11}_{BC}	N^{12}_{BC}
	Sector 2	N^{21}_{BA}	N^{22}_{BA}	N^{21}_{BB}	N^{22}_{BB}	N^{21}_{BC}	N^{22}_{BC}
REGION C	Sector 1	N^{11}_{CA}	N^{12}_{CA}	N^{11}_{CB}	N^{12}_{CB}	N^{11}_{CC}	N^{12}_{CC}
	Sector 2	N^{21}_{CA}	N^{22}_{CA}	N^{21}_{CB}	N^{22}_{CB}	N^{21}_{CC}	N^{22}_{CC}

Fig. 4.6.3b Variables for a 3-region, 2-sector MRIO model.

where $t^m{}_{ij}$ is the percentage of sector m product used in zone j (for whatever use) acquired from production zone i, known as inter-regional exchange or trade coefficient. Furthermore, it follows from construction:

$$t^m_{ij} = Z^{mn}_{ij} / K^{mn}_{ij} = Z^{mn'}_{ij} / K^{mn'}_{ij} = W^m_{ij} / Y^m_j$$

$$\Sigma_i \, t^m{}_{ij} = 1$$

The trade coefficients can be arranged in a matrix $\boldsymbol{T}$, known as exchange or trade matrix, of dimensions $(n_z \cdot n_s \times n_z \cdot n_s)$ in which for each pair of regions there is a diagonal matrix of dimensions equal to the number of sectors whose elements are the trade coefficients between the two regions, for the sector corresponding to the row and to the column. Fig. 4.6.3b reports an example of the matrix $\boldsymbol{T}$ for a 3-region, 2-sector system. Combining equations (4.6.2), (4.6.3) and (4.6.4) yields:

$$X^m{}_i + J^m{}_i = \Sigma_j \, \Sigma_n \, t^m{}_{ij} \cdot a^{mn}{}_j X^n{}_j + \Sigma_j \, t^m{}_{ij} \cdot Y^m{}_j \qquad (4.6.5)$$

which in vector terms can be expressed as:

$$\boldsymbol{X} + \boldsymbol{J} = \boldsymbol{TAX} + \boldsymbol{TY} \tag{4.6.6}$$

The model (4.6.6) is usually applied for the prediction of regional production for each sector, i.e., for the calculation of vector $\boldsymbol{X}$, starting from scenario hypotheses on the vector of final consumption $\boldsymbol{Y}$ and import $\boldsymbol{J}$. Once the vector $\boldsymbol{X}$ has been calculated, it is possible to estimate the O-D freight demand in quantity, as will be shown later.

The Multi-Regional Input-Output model[22] with constant coefficients assumes that the elements of matrices $\boldsymbol{A}$ and $\boldsymbol{T}$ are constant and known (equal, for example, to the respective present values). In this case, the solution of the linear equation system (4.6.6) can be expressed in closed form as:

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{TA})^{-1} \cdot (\boldsymbol{TY} - \boldsymbol{J}) \tag{4.6.7}$$

where $\boldsymbol{I}$ is the identity matrix of dimensions $(n_z \cdot n_s \times n_z \cdot n_s)$.

Fig. 4.6.4 gives a numerical example of the application of model (4.6.7) for a case of 3 regions and 2 sectors. Analysis of the results provides some general indications on the performances of MRIO models. If the value of the final demand of a zone increases, the values of production increase also in other zones. The example presents two scenarios. The second scenario assumes an increase in the final demand of region A ($Y^2_{iA} > Y^1_{iA}$) which causes an increase in production of the different sectors in the same region and in the other regions. Furthermore the increase of production in zone B is larger than that in zone C since the former has exchange coefficients with zone A larger than the latter, due for example to lower transportation costs. It can also be observed that since the increase of final demand in zone A is greater for sector 2 (+300) than for sector 1 (+200) and the production technology of sector 2 makes greater use of intermediate products of the same sector, the increase of production in sector 2 is greater than that of sector 1 in all regions.

MRIO models with variable coefficients. The application of the MRIO model with constant coefficients assumes the independence of exchange and technical coefficients from variations of some significant variables, such as level of production, relative prices and generalized transportation costs. These hypotheses are reasonable only for short-term forecasts. To overcome these shortcomings, various extensions of model (4.6.7) have been proposed. These basically consist of expressing the exchange coefficients (matrix $\boldsymbol{T}$) and/or the technical production coefficients (matrix $\boldsymbol{A}$) as functions of other transportation and economic variables. In this sense, they can be referred to as variable coefficient models.

In a first specification, known as a *MRIO model with elastic trade coefficients*, the coefficients t^m_{ij} are obtained with an explicit model that can be descriptive or a random utility model simulating the choice of supply zone. It is usually assumed for a number of reasons (dishomogeneity of products within the sectors, market

mechanisms differing from pure competition, omitted attributes, etc.) that supply acquisition is "dispersed" (probabilistic model), i.e. supply does not come exclusively from the zone(s) of minimum mean acquisition cost (deterministic model).

Technical coefficient matrix ***A*** (3·2x3·2)

		REGION A		REGION B		REGION C	
		Sector 1	sector 2	sector 1	sector 2	sector 1	Sector 2
REGION A	Sector 1	0.30	0.10	0.00	0.00	0.00	0.00
	Sector 2	0.20	0.40	0.00	0.00	0.00	0.00
REGION B	Sector 1	0.00	0.00	0.40	0.20	0.00	0.00
	Sector 2	0.00	0.00	0.30	0.70	0.00	0.00
REGION C	Sector 1	0.00	0.00	0.00	0.00	0.35	0.20
	Sector 2	0.00	0.00	0.00	0.00	0.25	0.40

Matrix of exchange or trade coefficients ***T*** (3·2x3·2)

		REGION A		REGION B		REGION C	
		Sector 1	Sector 2	Sector 1	Sector 2	Sector 1	Sector 2
REGION A	Sector 1	0.50	0.00	0.30	0.00	0.10	0.00
	Sector 2	0.00	0.40	0.00	0.35	0.00	0.15
REGION B	Sector 1	0.30	0.00	0.60	0.00	0.20	0.00
	Sector 2	0.00	0.35	0.00	0.50	0.00	0.25
REGION C	Sector 1	0.20	0.00	0.10	0.00	0.70	0.00
	Sector 2	0.00	0.25	0.00	0.15	0.00	0.60

Vectors of import ***J*** (3·2x1) *and of final consumption* ***Y*** (2 *hypotheses*)

	sector	*J*	*Y1*	*Y2*
REGION A	1	20	100	300
	2	30	200	500
REGION B	1	10	400	400
	2	10	200	200
REGION C	1	50	300	300
	2	30	300	300

Vector of sectorial production ***X*** (3·2x1) *for the* 2 *hypotheses* ***Y***

	Results		
	sector	*X1*	*X2*
REGION A	1	453	652
	2	659	969
REGION B	1	725	919
	2	877	1222
REGION C	1	534	675
	2	689	928

Fig. 4.6.4 Numerical example of a 3-region, 2-sector MRIO model.

The systematic utility of acquiring from zone i product m used in zone j, V^m_{ij}, is usually a function of several variables among which is the total production of sector m in zone i, X^m_i, and the average unit acquisition cost q^m_{ij}:

$$V^m_{ij} = V(X^m_i, q^m_{ij}) \tag{4.6.8}$$

In applications, acquisition percentages are usually simulated with a Multinomial Logit model:

$$t^m_{ij} = exp(V^m_{ij})/\Sigma_k exp(V^m_{kj}) \tag{4.6.9}$$

In general, therefore, the whole trade matrix is a function of the vector $\boldsymbol{X}$ and of the acquisition cost matrix $\boldsymbol{q}$.

$$\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{V}(\boldsymbol{X}, \boldsymbol{q})) \tag{4.6.10}$$

The interpretation of the attributes introduced into the specification of acquisition percentages requires further comment. The value of the total production of sector m in zone i, X^m_i, can be considered a proxy of supply variety. More properly, this attribute should be used through its logarithm ($ln\ X^m_i$) and considered as a size function (see section 4.3.2) expressing the unknown number of elementary choice alternatives. If there are other attributes, M^m_{kj}, correlated to the number of production units, the size function would assume the more general expression:

$$ln\left(X^m_i + \sum_k \gamma_k M^m_{ki}\right)$$

A non-behavioral interpretation of equations (4.6.8) and (4.6.9) simply assumes that the acquisition percentage of a certain zone is greater the lower the acquisition cost and the larger its production. This may be due to localization behavior in which production units set up near their supply and/or distribution markets.

In the most general case, the average unit acquisition cost q^m_{ij} can be expressed as a function of the average unit price (price index) of products m in i, p^m_i, and of the average unit transport cost of product m from i to j, c^m_{ij}:

$$q^m_{ij} = p^m_i + c^m_{ij} \tag{4.6.11}$$

The average unit transportation cost can in turn be expressed as a function of the generalized transportation costs of the different modes/services available between the two zones, either as a weighted average of these costs, or as an inclusive (EMPU) variable of a random utility mode/service choice model. For example, in the national model described so far, trade coefficients were simulated through a Multinomial Logit model, in which the sale prices p^m_i were assumed to have no influence (i.e. equal for all zones). The specification of the systematic utility adopted for this model is:

$$V^m_{ij} = \beta^m_1 C^m_{ij} + \beta^m_2 Region_{ij} + \beta^m_3 ln(X^m_j)$$

where:

C^m_{ij} is the logsum of transport costs derived from the mode choice model;
$Region_{ij}$ is the dummy variable for the same region, equal to 1 if $i = j$, 0 otherwise;
X^m_j is the total production of the region j in sector m.

The Multi-Regional Input-Output model with elastic trade coefficients can be expressed formally by substituting the expression (4.6.10) in the general equation (4.6.6):

$$\boldsymbol{X^*} + \boldsymbol{J} = \boldsymbol{T(X^*,q)AX^*} + \boldsymbol{T(X^*,q)Y} \tag{4.6.12}$$

From the previous equation it follows immediately that the production vector $\boldsymbol{X}$ can no longer be obtained as the solution of a system of linear equations (4.6.7), since the coefficients are non-linear functions of the unknown vector $\boldsymbol{X}$ through the expressions (4.6.10). The calculation of the vector $\boldsymbol{X^*}$ can therefore be traced to the solution of a fixed-point problem; the theoretical properties and solution algorithms of fixed-point problems are briefly described in Appendix A.

The model described can be further generalized in different ways depending on which variables are simulated (predicted) as endogenous variables. A model which could be defined *MRIO with elastic prices* introduces the mechanisms of average unit prices formation. Unit sale prices of product m in zone i, p^m_i depend on the average unit production cost of m in i, k^m_i and on the unit added value (labor, capital, profits, etc.) to production e^m_i. The former, in turn, depends on the average unit acquisition cost of intermediate goods and services, h, necessary for production of m, $\bar{q}^h_i$. In formal terms:

$$p^m_i = k^m_i + e^m_i \tag{4.6.13}$$

with:

$$k^m_i = \sum_h a^{hm}_i \, \bar{q}^h_i$$

Note that in (4.6.13), the technical coefficients a^{hm} are to be interpreted as the "quantity" of product h necessary to produce a unit of product m in zone i. The average unit acquisition cost of h in i can in turn be expressed as a weighted average of the unit acquisition costs from the different zones l "producing" h:

$$\bar{q}^h_i = \sum_l q^h_{li} t^h_{li} \tag{4.6.14}$$

From expression (4.6.11), (4.6.13) and (4.6.14) it can be deduced that the vector of unit acquisition costs depends on itself through prices, and on trade coefficients:

$$\boldsymbol{q^*} = f[\boldsymbol{q^*}, \boldsymbol{T(X)}, \ldots]$$

In this case an equilibrium configuration, $\boldsymbol{q}^*$, must be found for the vector $\boldsymbol{q}$. The problem (4.6.12) gets further complicated since $\boldsymbol{q}$ in this case also depends on the unknown vector $\boldsymbol{X}$.

The model can be further extended and generalized along several lines. One extension is to introduce productive capacity constraints for the different zones. In this case the price $p^m{}_i$, or rather the added value $e^m{}_i$, can be expressed as a function of the ratio between the production demand $X^m{}_i$ given by (4.6.2) and the production capacity in order to take into account mechanisms of rent formation. In other words, if the production of a sector required for intermediate and final uses exceeds the productive capacity of zone i, a supply-demand re-equilibration mechanism is triggered. This causes an increase of the sale prices $p^m{}_i$, therefore reducing the acquisition percentages from that zone (see equation (4.6.10)) until an equilibrium configuration between demand and production capacity is reached.

Another line of extension is to express the dependence on prices of other key variables, such as technical coefficients, imports and family consumption. For example, elements $a^{mn}{}_i$ of matrix $\boldsymbol{A}$ can be replaced by functions $a^{mn}{}_i(X^n{}_i, \boldsymbol{q}_i)$ which may depend on the total level of production of sector n in zone i, $X^n{}_i$ to take into account scale (dis)economies, and on the vector of average unit acquisition costs for intermediate factors, to take into account possible substitutions between the factors as functions of the relative acquisition costs. For (dis)economies of scale the quantity of product m necessary to produce a unit n diminishes (increases) as the total production of n increases. For substitution effects, the quantity of a product m whose acquisition cost is particularly high, used to produce a unit of n, can be reduced by using a greater quantity of another factor. In this type of model, added value factors, in particular labor, are usually explicitly included; also, in the vector of the final demand, household consumption is usually assumed dependent on the available income of families in each zone.

Value-quantity transformation of trades. Once the vector $\boldsymbol{X}$ of production for each sector and region has been calculated with one of the expressions (4.6.7) or (4.6.12), it is possible to compute the resulting exchange or trade matrix $\boldsymbol{N}$ whose elements $N^{nm}{}_{ij}$ represent the value of sector n produced in zone i consumed by sector m in zone j. The trade matrix $\boldsymbol{N}$ has dimensions $(n_z \cdot n_s \times n_z \cdot n_s)$ and is obtained by ordering blocks of dimensions $(n_s \times n_s)$ representing the monetary value of the products of each sector exchanged with each other sector for a given pair of production and consumption regions. Fig. 4.6.3 gives an example of the structure of the matrix $\boldsymbol{N}$ in the case of three regions and two sectors. Matrix $\boldsymbol{N}$ can be expressed as a function of the variables obtained by solving the Input-Output model or one of its generalizations such as:

$$\boldsymbol{N} = \boldsymbol{T}\,\boldsymbol{A}\,\boldsymbol{Dg}(\boldsymbol{X}) + \boldsymbol{T}\,\boldsymbol{Dg}(\boldsymbol{Y}) \qquad (4.6.15)$$

where the matrices $\boldsymbol{Dg}(\boldsymbol{X})$ and $\boldsymbol{Dg}(\boldsymbol{Y})$ are obtained by arranging the elements of the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively along the main diagonal of a square matrix with $(n_z \cdot n_s)$ rows and columns.

Finally, from matrix $\boldsymbol{N}$ it is possible to obtain the flows N^n_{ij} of goods produced in sector n (thus excluding service sectors) in zone i and consumed in zone j. These flows are expressed in monetary units and can be computed by adding up the values corresponding to all consumption sectors:

$$N^n_{ij} = \Sigma_m N^{nm}_{ij}$$

The last step is the transformation of the O-D matrices, N^n_{ij}, from values into physical quantities (tons) by goods typology (market segments). This transformation is normally conducted on the basis of coefficients estimated on the present situation, and then modified exogenously in forecasting scenarios[23]. The goods typologies, identified on the basis of shipment size and/or of manufacturing company, are closely linked to the structure and to the attributes of the mode choice models which will be dealt with in the next section.

In conclusion, for the prediction of freight transport demand, several models with different levels of complexity and different input data requirements are available. The most highly-structured formulations of such models[24] aim to simulate the entire economy from which goods exchange demand is deduced; such a level of generality, however, requires a considerable amount of data which may be unnecessary if the aim of the models is limited to the simulation of freight transport demand.

A further consideration concerns the interaction between macro-economic and transport models. In the formulations described above, generalized transport costs c^m_{ij} are assumed to be known, and these in turn depend on the production costs of carriers such as road and railways haulage companies. These costs depend on several different factors including the "objective" level-of-service variables for the various modes (travel times, congestion levels, etc.) as well as the carriers production structure (production functions). It is therefore possible, at least in principle, to introduce further feedback cycles and related equilibrium problems between generalized transport costs and goods (and passenger) flow on the various modal networks through mode and path choice models.

4.6.2. Freight mode choice models

Several formulations have been proposed for models simulating the distribution of freight demand between different transport modes and services. These models are derived from different approaches (descriptive, micro-economic, inventory, random utility). In the following, models based on the random utility paradigm will be described, since they are consistent with the general approach to demand modeling adopted in this volume, and many of the models proposed following other

approaches, can be extended and considered as generalizations of random utility models.

Random utility models applied to simulate freight modal choice can be divided into *aggregate* and *disaggregate*, on the basis of the data used for their specifications/calibration, and application. Aggregate models are based on data and attributes corresponding to aggregate freight flows between different zones with available transport modes. These models use mainly level-of-service attributes (e.g. average consignment times, average prices, etc.). Aggregate models, although simple to apply, have proved to have limited analysis capabilities since many important decision factors cannot be taken into account without a greater level of disaggregation.

For these reasons, disaggregate mode choice models have recently been studied more frequently. These typically refer to the random utility paradigm and can be divided in two types: *consignment models* simulating mode choice for individual consignments, and *logistic models* simulating a sequence of logistic choices including the size and frequency of consignments, as well as the transportation mode.

Consignment mode choice models are more frequently used in applications. They usually have a functional form that belongs to the Logit family, most often of the Multinomial Logit type although Hierarchical Logit models have also been proposed in several applications. Choice alternatives typically correspond to the transport modes available for a given consignment (truck, train, ship, air) and often different services are also distinguished (e.g., conventional railway or combined road/railway, etc.). The level-of-service attributes normally used are consignment time, cost, reliability, etc. Other attributes usually included in specifications correspond to characteristics of the consignment (e.g., size, goods typology, frequency) and of the firm (e.g. annual invoicing, availability of own trucks or availability of railway sidings). Fig. 4.6.5 shows an example of a consignment mode choice model calibrated for the national model.

Logistic mode choice models are newer and so far have found few applications in spite of their theoretical interest and their usefulness for evaluating innovative supply combinations (logistic + transport services). These models simulate mode choice in the context of the logistic decisions of the firm determining the transport mode which, depending on the case, may be the selling or purchasing firm. In particular, it is assumed that the choice of transport mode depends on the *logistic cost* connected with its use, which in turn is made up of different components such as:

- costs associated with orders management;
- costs of transport (prices required for the transport service);
- costs associated with loss and damage;
- costs of capital locked up during transport;
- costs of carrying inventory;
- costs connected with the non-availability or delayed arrival of equipment for transport;

- costs of unreliability (early or delayed arrival and related costs of longer storage or locking up of larger supplies).

Alternatives: Train, Road, Combined Rail+Road

$$V_{train} = \beta_{T_t} T_t + \beta_{Mc} Mc_t + \beta_{p>30} \cdot p > 30 + \beta_{HVG} \cdot HVG + \beta_{train} \cdot TRAIN$$
$$V_{road} = \beta_{T_r} T_r + \beta_{Mc} Mc_r + \beta_{PSH} \cdot PSH$$
$$V_{combined} = \beta_{T_c} T_c + \beta_{Mc} Mc_c + \beta_{COMB} \cdot COMB$$

T_t	= "train" travel time
T_r	= "road" travel time
T_c	= "combined" travel time
Mc_t	= "train" monetary cost
Mc_r	= "road" monetary cost
Mc_c	= "combined" monetary cost
p>30	= Dummy variable: 1 if the shipment weights more than 30*t*, 0 otherwise
PSH	= Dummy variable: 1 if goods carried are perishable, 0 otherwise
HVG	= Dummy variable: 1 for of high value goods, 0 otherwise
TRAIN, *COMB*	= Alternative Specific Attributes (ASA)

	T_t	T_r	T_c	*Mc*	p>30	*PSH*	*HVG*	*TRAIN*	*COMB*
	-0.06	-0.15	-0.12	-1.47	1.20	0.86	-0.64	0.29	-3.34
t	*-1.7*	*-2.2*	*-2.0*	*-3.2*	*0.6*	*1.1*	*-1.2*	*0.5*	*-2.5*

Fig. 4.6.5 Example of freight "consignment" mode choice model.

Logistic costs depend on several factors such as the total (annual) quantity of consignments over a given commercial relation, the average frequency and size of the consignments, and the value of the goods. Furthermore, they depend on the characteristics of the service offered by the different modes such as price, reliability of consignment times, and the possibility of theft and damages. Direct information on all of the components of the logistic cost is very difficult to obtain, so it is assumed that the systematic utility function for each mode j is a combination of variables X^i_{jk} linked to the logistic cost items of a certain commercial relation i and that the coefficients β_k are the unknown cost factors. In any case, a great deal of information is required to specify and calibrate these models and their use, at the time, is mostly limited to the analysis of the factors influencing mode choice rather than to large-scale applications.

Reference Notes

The literature on transportation demand models is very broad and covers a period of more than forty years.

The first "partial share" demand model systems were formulated in the 50's and 60's although with time they have undergone a number of developments, both

formal and interpretive. A treatment of the "traditional" system, essentially descriptive, can be found in the books by Wilson (1974) and Hutchinson (1974).

Since the mid-70's, several model systems have been proposed for trip simulation, inspired by the theory of random utility. Some examples can be found in the books by Domencich and McFadden (1975), Richards and Ben Akiva (1975), Manheim (1979), Ben Akiva and Lerman (1985), Ortuzar and Willumsen (1994). The systems of random utility models proposed in the literature are mainly based on the factorialization of Logit and Hierarchical Logit models. The general formulation of systems of partial share models based on different random utility models integrated through EMPU variables proposed in paragraph 4.2.1 is original.

International literature proposes several theoretical contributions and applications of trip frequency, distribution, mode and path choice models both at urban and at extra-urban level.

Among the first examples of trip emission models of the category analysis type, the work of Oi and Shuldinher (1972) should be mentioned; an example of behavioral models of trip frequency at the urban level is contained in Biggiero (1991), and at the inter-urban level in Cascetta, Nuzzolo and Biggiero (1995).

Distribution models with size functions were proposed by Richards and Ben Akiva (1975), Koppelman and Hauser (1978), Kitamura et al. (1979); a summary can be found in Ben Akiva and Lerman (1985). References to descriptive or gravitational distribution models can be found in the above-mentioned text by Wilson (1974). An example of behavioral destination choice model in urban areas with explicit choice set simulation is contained in Cascetta and Papola (2000).
Modal split models of the Logit or Nested Logit type reported in the literature are extremely numerous; many examples are given in the mentioned books by Ben Akiva and Lerman (1985) and Willumsen and Ortuzar (1994).

A systematic analysis of the different hypotheses at the basis of path choice models is contained in Cascetta (1995). Path choice models for road networks calibrated on empirical data are not numerous, among them the works of Ben Akiva et al. (1984), Cascetta, Nuzzolo and Biggiero (1995), Russo and Vitetta (1995) can be mentioned. The specification of the Probit path choice model is described in Sheffi (1985). The C-Logit model is described in Cascetta, Nuzzolo, Russo and Vitetta (1996).

As for path choice models on public transport networks, the interpretation of pre-trip/en-route behaviors is described in Cascetta and Nuzzolo (1986), the concept of travel strategy is formulated in Spiess and Florian (1989), the representation as network hyperpath was proposed by Nguyen and Pallottino (1986).

The system of interurban travel demand simulation models described in section 4.3.5 was calibrated for the Italian National Modal System SIMPT and is described in Cascetta et al. (1995).

The literature describes several trip chaining (journey) demand models; an analysis with a bibliographical commentary is in Ben Akiva, Bowman and Gopinath (1995). Some trip-chaining models based on the concept of primary destination (activity) are described in Antonisse, Daly and Gunn (1986) and Algers, Daly and

Widlert (1993). The models system described in section 4.4 is based on the work of Cascetta et al. (1994).

The models relating freight demand to production and consumption in the economic system can be divided into two groups: Spatial Price Equilibrium (SPE) and Sectorial Interdependences models. A systematic description of the contributions relative to the former group is in Friesz et al. (1983) and in the books by Harker (1985) and (1987).

Multiregional Input-Output models with constant coefficients are generalizations of the model proposed by Isard (1951) and later developed for freight transport demand; see Leontief and Costa (1987) and Costa and Roson (1988). The MRIO model with elastic trade coefficients applied for the Italian case is described in Cascetta et al. (1996). The generalization which includes in the formulation the prices equilibrium mechanism is original. For a classification of Computable General Equilibrium (CGE) models reference can be made to Bergman (1990).

The bibliography on freight modal split models is quite substantial. An analysis of the factors influencing the behavior of different operators can be found in the volume by Bayliss (1988), an analysis and classification of the different modal split models is in Winston (1983), examples of disaggregate consignment models calibrated in Italy are described in Nuzzolo and Russo (1995). For a description of the theoretical assumptions of logistic random utility models with some empirical results reference can be made to Modenese and Vieira (1992).

Notes

(1) Demand models (4.1.1) are typically obtained through the integration of several sub-models. In this respect it would be more appropriate to speak of a system of demand models. The definition of demand model corresponds to the micro-economic concept of aggregate demand function for transport services.

(2) Note that the vector $\boldsymbol{T}$ includes individual level of service or performance attributes, while the generalized costs are combinations of level of service attributes; the homogenization coefficients are among the parameters of demand models.

(3) All the models described in this chapter depend on coefficients or parameters, which, for the time being, will be assumed known. Calibration of the models, i.e the estimation of unknown parameters, will be dealt with in detail in Chapter 8.

(4) From now on reference will be made to passenger transport demand, even though many of the concepts introduced can be extended to freight transport demand models, which will be dealt with in section 4.6.

(5) Differences between behavioral and descriptive models are increasingly less important. In fact, more and more often functional forms such as Logit and Hierarchical Logit deriving from the theory of random utility are used to simulate aspects of demand which have no direct behavioral interpretation in terms of the decision-maker's choice. From this point of view it would be more appropriate to classify the models on the basis of their functional form, distinguishing between models that can or cannot be derived from the theory of random utility.

(6) It should also be noted that the level of aggregation might be different in the phases of calibration and application of the model. In other words it is possible, and even convenient in some cases, to use individual information for the specification and calibration, as will be seen in Chapter 8, while in

applications average values of zone and users category can be used. This corresponds to the application of the aggregation techniques "by average user" or "by category" described in section 3.7.

(7) A trip is sometimes identified by a single purpose, e.g. work, study, etc. This practice may cause confusion; it would be more correct to define the purpose s of a single trip by a pair of purposes, i.e. the activities carried out at the origin and at the destination. In this way, the purpose house-to-work (H-W) is different from the purpose work-to-work (W-W). Trips for which the purpose "home" appears in origin or destination are often indicated as "home-based", others as "secondary". The characterization of a trip with a pair of purposes also allows the more precise identification of the variables of the activities system to which reference can be made.

(8) Note that in this case the random residual τ^*_d is not distributed like the normal multivariate r.v. $\tau_{m'd}$ since the Normal r.v. is not stable with respect to maximization (the variable τ^*_d has no known probability law).

(9) It is worth noting that the whole model is still monotone increasing with respect to systematic utilities as proved in subsection 5.6.

(10) From the analysis of the coefficients different indications on the socio-economic factors influencing non-systematic mobility (i.e. related to purposes different from commuting and study) in urban areas can be deduced. For example, the results reported in Fig. 4.3.1 suggest that the frequency of activities (and trips) increases with income level. The accessibility of the residence zone with respect to the location of commercial activities increases the trip frequency for shopping, but is not significant for business and personal services trips. There is a greater tendency for women and unemployed persons to undertake trips; young people tend to have less mobility, in the time period considered, especially for shopping; there is an effect of substitution with other members of the family for shopping (positive coefficient for the TOF variable), while there is an effect of complementarity (accompanying) for the other purposes (negative coefficient of TOF). Carrying out other activities (coefficient of the TOP variable) reduces the time available to engage in the activity (trip purpose) considered and so on. Note, also, that the coefficient of accessibility corresponds to the parameter δ_x in equation (4.2.12a) and turns out to be included in the interval (0,1), coherently with the behavioral interpretation of the model.

(11) The aggregation technique of sample enumeration, described in section 3.7, should therefore be used for more sophisticated specifications of the models.

(12) Gravitational models derive their name from the formal similarity of their earlier specifications with the Newton's law of universal gravitation. Subsequently simply and doubly constrained gravity models were derived from Entropy maximization principles. In this approach a measure of the number of possible micro-states (i.e. individual trips between each origin-destination pair) is expressed by the entropy measure of a given trip distribution; the entropy function is maximized under some constraints expressing the total number of trips leaving and/or reaching each zone and the total transportation cost (distance) spent in the system. The resulting maximum entropy distribution models are usually referred to as simply and doubly constrained gravity models. These model, though still quite popular in applications, do not allow the flexibility of the whole set of random utility models, regardless of their behavioral interpretation, nor the possibility of introducing other attributes explaining attractiveness and perception of alternative destinations. On the other hand it should be said that relatively little research effort has been dedicated to the analysis of behaviorally more complex models of destination choice. As a matter of fact, the assumption of i.i.d. residuals underlying the Multinomial Logit structure is questionable in the case of spatially contiguous traffic zones. In this case Cross-Nested Logit or Probit models should be used. Also models or variables explaining different levels of destination perception (choice set modeling) should be included.

(13) Pure pre-trip behavior assumes that users don't modify the route chosen at the beginning of the trip. In reality there are situations in which the user modifies his/her route by adapting to conditions encountered during the journey even for continuous service modes (e.g. accidents and unexpected jams). This type of behavior is even more prevalent when there are information technologies (variable message

signs, radio news, on-board computers) providing information on the state of the network or suggest the route to take in real time. Path choice models for continuous service networks which take account "mixed" behavior, are however still at the research stage and will not be dealt with here. Furthermore static assignment models are meant to simulate recurrent congestion, thus path choice models included in static assignment can be assumed to be based on such recurrent conditions and thus rule out accidents or other exceptional events.

(14) Another advantage of explicit path enumeration includes the possibility of solving certain algorithmic problems in the application of a Monte Carlo algorithm for the calculation of choice probabilities for a Multinomial Probit model, as will be seen in Chapter 7.

(15) More sophisticated specifications introduce also socio-economic attributes of the driver such gender, income etc.

(16) When the generalized cost depends on a single attribute (such as travel time in urban networks), this is multiplied by a coefficient β of homogenization in utility terms.

(17) Path choice for regular, low frequency scheduled services with explicit run representation is usually assumed to be completely pre-trip and the models simulating it are analogous to those described for road networks. In this case, however, the choice alternatives are the single runs or sequences of runs which can be represented as paths on the diachronic network. This point will be dealt with extensively in Chapter 6.

(18) More complex rules of en-route behavior have been proposed. In these cases the user boards the vehicle on its arrival at a stop, or waits for the vehicle of another line, comparing the expected value of the cost of the different options on the basis of the information available at that moment. These models of en-route behavior require a great deal of information and get close to the micro-simulation of network journeys; for these reasons, they are not used for demand assignment to large scale networks. Models of this type will be described in Chapter 6 for irregular scheduled services.

(19) Diversion nodes are those in which the outgoing links represent the connections of the stop with the lines serving that stop, as defined in Chapter 2 and represented graphically in Fig. 4.3.10. For a formal definition of the hyperpath in terms of graph variables, see section 5.5.

(20) It is assumed that h_I is defined such that the probability of undertaking more than one journey for the same purpose in the same time period is negligible.

(21) Note that the definition of a user class as individuals who have the same demand models (alternatives, parameters and attributes) is linked to what models, and therefore to what choice dimensions, reference is made. In particular, in Chapter 5 the classes will be defined with reference to path choice models. Given the reduced number of attributes in these models, it may happen that fewer classes are used for assignment than for other choice dimensions; the former can be obtained by aggregation of the latter. This is particularly true in the case of models specified at individual level where the individuals can be aggregated to obtain O-D trips matrices belonging to a given user class for the assignment model.

(22) Input-Output models are said to be "demand-driven" since, as follows from expression (4.6.7), the production vector $\boldsymbol{X}$ is obtained starting from vectors of final demand $\boldsymbol{Y}$ and the imports $\boldsymbol{J}$. It is thus assumed that the supply productive capacity adapts itself to the production levels required by the demand.

(23) Value/quantity transformation coefficients can differ significantly from unit market prices since they should capture the differences between physical goods movements and commercial transactions. For example a single commercial transaction may correspond to several good movements due to intermediate stockage locations and so on. Given the increasing relevance of freight logistics on transport demand, value/quantity transformation coefficients should be explicitly modeled as functions of relevant variables of the logistic cycle of the industrial sector they refer to.

(24) Models of this type are known in the literature as *Computable General Equilibrium (CGE)* models.

5 MODELS FOR TRAFFIC ASSIGNMENT TO TRANSPORTATION NETWORKS(°)

5.1. Introduction

Models for traffic assignment to transportation networks simulate how demand and supply interact in transportation systems. These models allow the calculation of performance measures and user flows for each supply element (network link), resulting from origin-destination demand flows, path choice behavior, and the reciprocal interactions between supply and demand. Assignment models combine the supply and demand models described in the previous chapters; for this reason they are also referred to as demand-supply interaction models. In fact, as seen in Chapter 4, path choices and flows depend on path generalized costs, futhermore demand flows are generally influenced by path costs in choice dimensions such as mode and destination. Also, as seen in Chapter 2, link and path performance measures and costs may depend on flows due to congestion. There is therefore a circular dependence between demand, flows, and costs, which is represented in assignment models as can be seen in Fig. 5.1.1.

Assignment models play a central role in developing a complete model for a transportation system since the results of such models describe the state of the system, or the "average" state and its variation. Assignment results, in turn, are the inputs for the design and/or evaluation of transportation projects. Several classes of assignment models have been built on the basis of varying assumptions regarding the many components, including demand, supply, and the approach used for studying supply and demand interactions. These hypotheses determine some classification factors of assignment models, which will be defined in the following to introduce a general taxonomy for such models (see fig. 5.1.2 below).

(°) Giulio Erberto Cantarella is co-author of this Chapter.

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

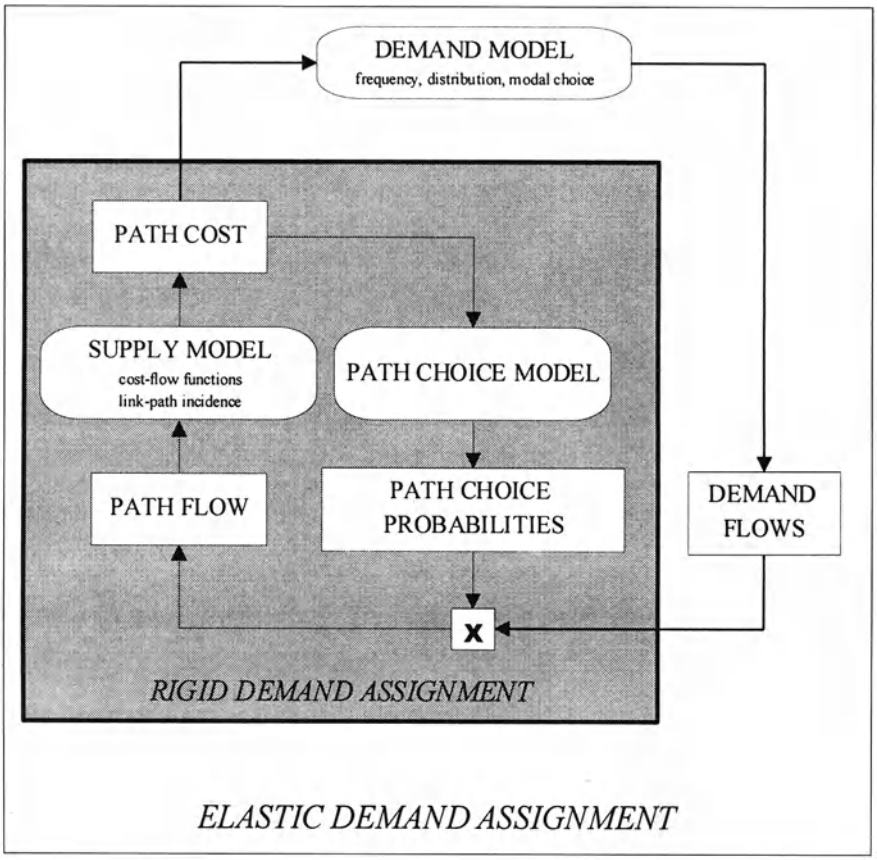


Fig. 5.1.1. – Schematic rappresentation of assignment models.

Assignment models can be classified with respect to *supply characteristics*, i.e. the type of transportation services available and the dependence of link performance variables on link flows (congestion). In particular, transportation services can be classified as *continuous* or *scheduled*, as introduced in Chapters 1 and 2. The dependence of link performance variables on flows is the other primary classification factor with respect to supply. When link and costs are independent of flows, i.e. congestion effects are negligible, *fixed cost* or *uncongested networks* (*UN*) assignment models result. On the other hand, if link costs depend on flows, *variable costs* or *congested network* assignment models are obtained.

Another set of classification factors relates to the *travel demand assumptions.* Assignment models can be distinguished with respect to the hypotheses regarding path choice behavior presented in Section 4.2.5. Generally, path choice may result from a sequence of decisions made at different times during a trip; in this case, we have *mixed pre-trip/en-route* behavior. In the pre-trip choice stage, conducted before

starting the journey, single paths or strategies for en-route choice among paths (hyperpaths) are considered as alternatives. In the latter case, the path actually followed during a particular trip is the result of en-route decisions made during the journey in response to unpredictable events. In the case of *fully pre-trip* behavior, single paths are the alternatives considered in the pre-trip choice stage. In both cases, pre-trip choice takes into account cost attributes corresponding to network characteristics. For users who move between an origin-destination pair, the pre-trip path choice model expresses the probability of choosing single paths or hyperpaths. Models based on random utility theory are typically used to simulate these choices. In particular, *deterministic (D)* choice models assume that the perceived utility of a path is deterministic, and all users choose a maximum average utility (minimum average cost) alternative. On the other hand, *probabilistic or stochastic (S)* choice models assume that the perceived utility of a path is a random variable, and that the users may choose any alternative, as described in Section 4.2.5.

A further classification factor related to demand is the dependence of O-D flows on path performance measures and costs. *Rigid demand* assignment models assume that demand flows are independent of cost variations due to network congestion. On the other hand *elastic demand* models assume that demand flows depend on congestion costs; demand flows are therefore a function of path costs resulting from congestion, as well as activity system attributes. Demand can be assumed elastic only on some dimensions; for example, the total O-D matrix can be assumed to be cost-independent (frequency and destination choices are not influenced by cost variations), while mode choice varies with link costs; in this way, *multi-mode assignment* models are obtained. Obviously, from a practical viewpoint, demand elasticity is relevant only for congested networks where costs depend on flows.

With respect to demand segmentation, assignment models are called *multi-user class* if users are subdivided into several classes, and all users in a class share the same choice model (attributes, coefficients and functional form). In this way, it is possible use different choice models for different trip purposes or user socio-economic categories (income, etc.). In the case of road systems, it is possible to distinguish different vehicle types (motorcycles, cars, commercial vehicles, etc.). *Single user class* assignment is a special case where all users share the same choice model and are distinguished by O-D pairs only.

Transportation systems can be represented under two different assumptions on the intra-period variability of variables. Consistent with the hypotheses presented in Chapter 1 regarding the definition of a transportation system, variations of demand and/or supply within the reference period (e.g. the morning peak-period) are not considered in this chapter. The assignment models presented are thus *intra-period* or *within-day static*. This hypothesis is realistic only if transportation demand and supply characteristics can be safely assumed constant over a reference period of sufficient length with respect to the journey times of the system. Otherwise, *intra-period* (or *within-day*) *dynamic* assignment models should be adopted, which require extensions of both the demand and, even more so, supply models. Within-day

dynamic assignment models can also be distinguished according to all the criteria discussed in this section and will be addressed in Chapter 6.

Another fundamental classification factor for assignment models is the approach used for studying supply and demand interactions. In particular, *user equilibrium assignment*[1] models represent equilibrium configurations of the system, i.e. configurations in which demand, path, and link flows are mutually consistent with the costs that they induce. From a mathematical point of view, equilibrium assignment can be defined as the problem of finding a flow vector that reproduces itself on the basis of the correspondence defined by the supply and demand models. This problem can be formulated with fixed-point models, or with variational inequality or optimization models, as will be shown in the following sections.

The alternative approach leads to *inter-periodal (or day-to-day) dynamic process* assignment models. In this case it is assumed that the system might evolve over time (i.e. successive reference periods), through possibly different feasible states, due to several causes such as the variability of the number of users undertaking trips, path choices, supply performance measures, etc. One of the mechanisms of the evolution from one state to another is the flow - cost dependence. In fact, if in a reference period the system is in a given state – demand, path and link flows and costs – this state may be not internally consistent and may cause a change toward a different state in the following reference periods. Dynamic process assignment models explicitly simulate the evolution of the system state based on the mechanisms underlying path choice and information acquisition, which in turn determine user choices in successive reference periods. By analogy, equilibrium assignment is also known as day-to-day static assignment. *Dynamic process* models can be further distinguished as *deterministic* or *stochastic* depending on whether the system state is modeled using deterministic or stochastic (random) variables.

Figure 5.1.2 reports the different classification factors introduced. Although no such complete taxonomy is used in the technical literature, the identification of assignment models with the full set of factors in a sequence is a useful exercise, as it clearly identifies the assumptions underlying any particular model.

In the following sections, models obtained by combining the above-described elements are presented in increasing order of generality (and complexity). Section 5.2 reviews the main definitions and hypotheses adopted in the development of supply and demand models assuming a single user class, fully pre-trip path choice, and rigid demand.

Then, under these hypotheses, section 5.3 describes uncongested network assignment models and Section 5.4 congested network equilibrium assignment models. Extensions to mixed pre-trip/en-route path choice behavior are described in section 5.5, assignment with elastic demand and/or multi-modal systems is dealt with in section 5.6, and assignment with users belonging to different classes (multi-user class assignment) in section 5.7. Section 5.8 presents some general considerations about dynamic process assignment, which is still mainly at a research level. Section 5.9 presents some comments related to the application of assignment models to real systems.

Supply factors	
Type of service	Continuous Scheduled
Congestion effects	Uncongested Networks Congested Networks
Demand factors	
Demand Segmentation	Single User class Multiple User classes
Demand Elasticity	Rigid Demand Elastic Demand
Path Choice Behavior	Fully pre-trip Pre-trip/en-route
Path Choice Model	Deterministic Probabilistic
System approach factors	
Intra-periodal Variability	Intra-period Static Intra-period Dynamic
Demand-supply interaction	User equilibrium Deterministic Dynamic Process Stochastic Dynamic Process

Fig. 5.1.2. Classification factors of assignment models.

Extensions of supply, demand, and demand/supply interaction models to intra-periodal dynamic systems with continuous scheduled services will be discussed extensively in Chapter 6. The algorithms for solving assignment models, that is for the calculation of resulting link performance measures and flows, will be considered in Chapter 7.

5.2. Definitions, assumptions, and basic equations

This section summarizes the definitions and assumptions underlying the demand and supply models discussed in Chapters 2 and 4 respectively. A single mode is considered (*single-mode assignment*), and it is assumed that the O-D demand flows are known and independent of the congested link costs (*rigid demand assignment*). It follows that path choice is the only choice dimension explicitly simulated. Users are considered to be homogeneous, that is they share common behavioral and cost characteristics regardless of trip purpose and differ only by origin and destination (*single-users class assignment*). Also, path choice is considered completely pre-trip (as is often the case for road transport systems).

The symbols and definitions, already introduced in Chapters 2 and 4, are repeated below for reader convenience (with reference to a time band h and a mode m not explicitly indicated for the sake of simplicity).Let

o be a node (zone) origin of a trip ;
d be a node (zone) destination of a trip ;
od be an origin-destination pair;
K_{od} be the set of paths for od pair; a path k is univocally associated with od such that $k \in K_{od}$; the set K_{od} is not empty if at least one path connecting o and d is available, it is finite if only elementary (say loopless) paths are considered;
Δ_{od} be the link-path incidence matrix for od pair;
Δ be the overall link-path incidence matrix obtained by placing the blocks corresponding to each od pair side-by-side.

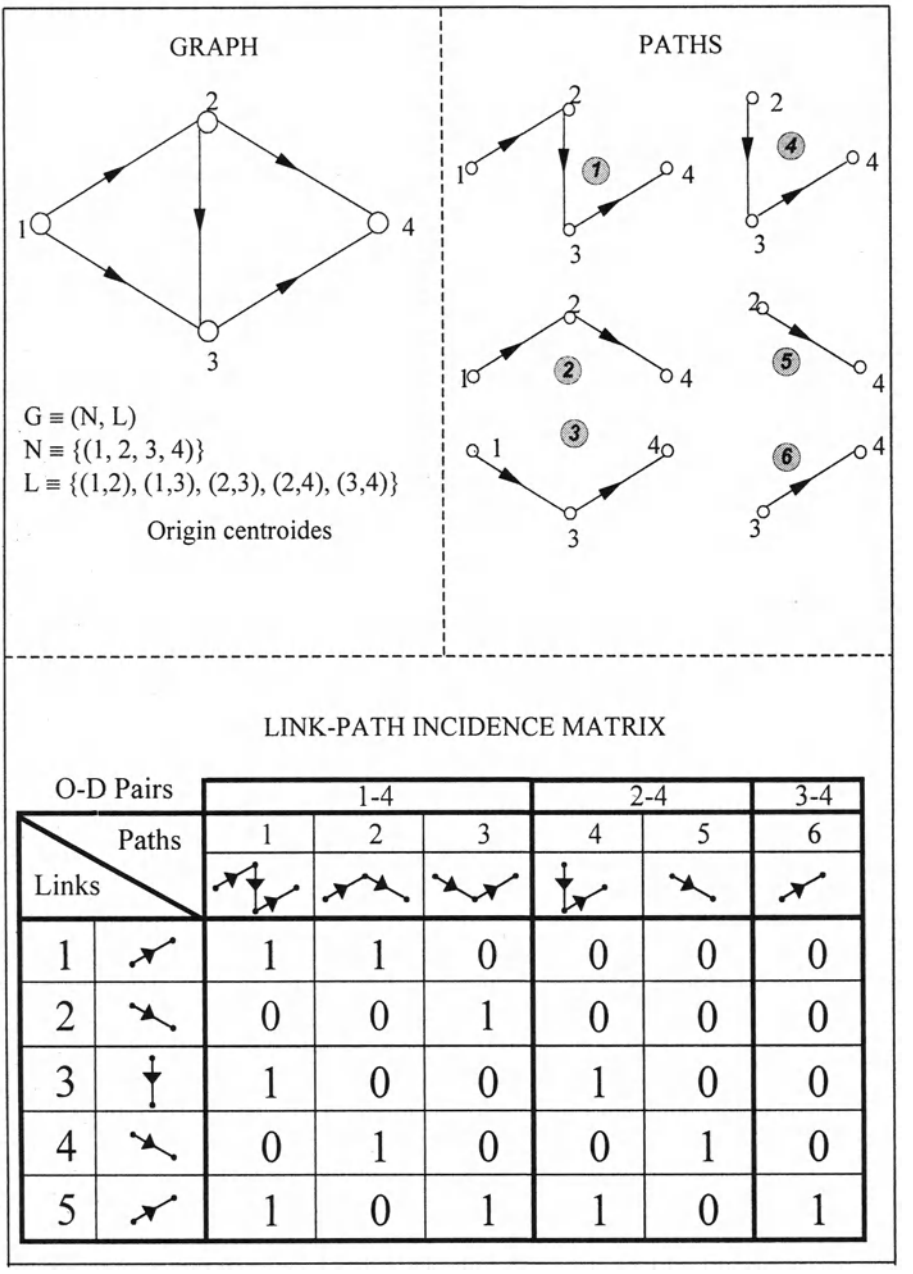
5.2.1. Supply model

Transportation supply is simulated with a (congested) network model, as described in Chapter 2. To each link a a (generalized) cost c_a is associated, measured in units homogeneous to the utility through appropriate homogenization coefficients. Throughout this chapter, it will be assumed that path (dis)utility function are linear with respect to path performance attributes. Furthermore, each path k is associated with a *path cost*[2], g_k, consisting of two types of cost attributes:

- *Link-wise additive (or generic) path costs* that are obtained by summing up the corresponding link costs, independent of the O-D pair and/or of the path; these costs may depend on the link flows in the case of congested networks;

- *Link-wise non-additive (or specific) path costs* that are specific to the path and/or to the O-D pair, since they cannot be defined by summing corresponding generic link costs. In the following analysis, these costs are assumed to be independent of congestion. Therefore, path costs that are simultaneously non-additive and dependent on congestion are not considered. Let:

$\boldsymbol{c}$ be the link cost vector with entries, c_a;
$\boldsymbol{g}_{od}^{ADD}$ be the vector of additive path costs g_k^{ADD} for the users of the od pair, $k \in K_{od}$;
$\boldsymbol{g}^{ADD}$ be the overall vector of additive path costs, consisting of the vectors of additive path costs $\boldsymbol{g}_{od}^{ADD}$ corresponding to all od pairs;
$\boldsymbol{g}_{od}^{NA}$ be the vector of non additive costs g_k^{NA} for the users of the od pair, $k \in K_{od}$;
$\boldsymbol{g}^{NA}$ be the overall vector of non additive path costs consisting of the vectors of non additive path costs $\boldsymbol{g}_{od}^{NA}$ corresponding to all od pairs;
$\boldsymbol{g}_{od}$ be the vector of total path costs g_k for the users of the od pair, $k \in K_{od}$;
$\boldsymbol{g}$ be the overall vector of the total path costs, consisting of the vectors of the total path costs $\boldsymbol{g}_{od}$ corresponding to all od pairs.



LINK-PATH INCIDENCE MATRIX

O-D Pairs		1-4			2-4		3-4
Links \ Paths		1	2	3	4	5	6
1		1	1	0	0	0	0
2		0	0	1	0	0	0
3		1	0	0	1	0	0
4		0	1	0	0	1	0
5		1	0	1	1	0	1

Fig. 5.2.1 – Example of a graph and its link path incidence matrix

The relationship between link costs and path costs is given for each *od* by the following equations (see Figure 5.2.2):

$$\boldsymbol{g}_{od}^{ADD} = \boldsymbol{\Delta}_{od}^{T}\, \boldsymbol{c} \qquad \forall od$$

$$\boldsymbol{g}_{od} = \boldsymbol{g}_{od}^{ADD} + \boldsymbol{g}_{od}^{NA} \quad \boldsymbol{\Delta}_{od}^{T}\, \boldsymbol{c} + \boldsymbol{g}_{od}^{NA} \qquad \forall od \tag{5.2.1a}$$

$$\boldsymbol{g} = \boldsymbol{g}^{ADD} + \boldsymbol{g}^{NA} = \boldsymbol{\Delta}^{T}\, \boldsymbol{c} + \boldsymbol{g}^{NA} \tag{5.2.1b}$$

$$\begin{bmatrix} 6 \\ 4 \\ 2 \\ \hline 4 \\ 2 \\ \hline 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \hline 0 \\ 0 \\ \hline 0 \end{bmatrix}$$

$\boldsymbol{g} = \quad \boldsymbol{\Delta}^{T} \quad \cdot \boldsymbol{c} \quad + \quad \boldsymbol{g}^{NA}$

Fig. 5.2.2 Example of the relationship between link costs and path costs (non additive costs are null for the sake of simplicity).

A flow f_l is associated with each link l. Link flows are measured in units homogeneous to demand flows. Let:

$\boldsymbol{f}$ be the link flow vector, with entries f_l.

In the case of congested networks, as described in Chapter 2, link costs depend on link flows through the following cost functions:

$$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f}) \tag{5.2.2}$$

In turn, link flows depend on the flow associated with each path and measured in units homogeneous to demand flows, through the *Network Flow Propagation* model. In particular, path flows for each *od* pair induce the corresponding link flows by *od* pair through the link-path incidence matrix. Furthermore, assuming that the demand flow for each O-D pair is expressed in consistent units, the total flow on a link is the sum of the flows induced by all O-D pairs. Let:

$\boldsymbol{h}_{od}$ be the path flow vector for the users of the *od* pair, the elements of which are the flows h_k for any index k in set K_{od};

$\boldsymbol{h}$ be the overall vector of path flows, consisting of the vectors of path flows $\boldsymbol{h}_{od}$ corresponding to each *od* pair;

$\boldsymbol{f}^{od}$ be the vector of link flows, f_l^{od}, corresponding to the *od* pair trips, measured in units consistent with the demand flows.

The relationship between link flows and path flows is expressed by the following equations (Fig. 5.2.3):

$$\boldsymbol{f}^{od} = \Delta_{od}\boldsymbol{h}_{od} \quad \forall od$$

from which

$$\boldsymbol{f} = \Sigma_{od}\boldsymbol{f}^{od} = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od} \tag{5.2.3a}$$

$$\boldsymbol{f} = \Delta\boldsymbol{h} \tag{5.2.3b}$$

$$\boldsymbol{f} = \quad \Delta \quad \cdot \boldsymbol{h}$$

$$\begin{bmatrix} 335 \\ 665 \\ 494 \\ 1341 \\ 1959 \end{bmatrix} = \left[\begin{array}{cc|c|cc|c} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} 90 \\ 245 \\ \underline{665} \\ 404 \\ \underline{1096} \\ 800 \end{bmatrix}$$

Fig. 5.2.3 – Example of the relationship between link flows and path flows.

The whole supply model is defined by equations (5.2.1-3), which combined together express the relationship between path costs and path flows already introduced in Chapter 2:

$$\boldsymbol{g}_{od} = \Delta_{od}^{T}\,\boldsymbol{c}(\Sigma_{od}\Delta_{od}\boldsymbol{h}_d) + \boldsymbol{g}_{od}^{NA} \quad \forall od \tag{5.2.4a}$$

$$\boldsymbol{g} = \Delta^{T}\boldsymbol{c}(\Delta\boldsymbol{h}) + \boldsymbol{g}^{NA} \tag{5.2.4b}$$

If cost functions are continuous with continuous first derivatives, the supply model is a continuous function with continuous first derivatives with respect to path flows.

5.2.2. Demand model

As stated earlier, for the time being it is assumed that demand flows are known and independent of cost variations; thus the path is the only choice dimension explicitly simulated. It is also assumed that the demand flows for different O-D pairs are expressed in mutually consistent units. In particular, for individual modes such as

car, they are measured in vehicles or drivers per unit of time, while in the case of public (scheduled) transport modes they are usually expressed in terms of passengers per unit of time. Let

$d_{od} \geq 0$ be the demand flow for the *od* pair, defined by the elements of the O-D matrix (corresponding to the purpose, the mode, and the time band to be analyzed);

$\boldsymbol{d}$ the demand vector, whose components are the demand values d_{od} for each O-D pair .

Path choice behavior is simulated with random utility models, assuming that a component of the systematic utility is equal to the opposite of the path generalized cost[3]: (Section 4.2.5):

$$\boldsymbol{V}_{od} = -\boldsymbol{g}_{od} + \boldsymbol{V}^{\circ}_{od} \quad \forall od \tag{5.2.5}$$

where:

$\boldsymbol{V}_{od}$ is a vector with elements consisting of the systematic utilities V_k of paths $k \in K_{od}$ for the users of the *od* pair;

$\boldsymbol{V}^{\circ}_{od}$ is a vector with elements consisting of the part of systematic utility depending on any other attributes which cannot be included in path costs (such as the users' socio-economic attributes), omitted in following sections for simplicity of notation.

Path choice probabilities depend on the systematic utilities of the available paths through the path choice model. Let:

$p_{od,k} = p[k/od]$ be the probability that a user, for a trip from origin *o* to destination *d* (without explicit indication of purpose, time band, and mode) will use the path *k*;

$\boldsymbol{p}_{od}$ be the vector of path choice probabilities for users of the *od* pair, whose elements are the probabilities $p_{od,k}$ with index *k* in set K_{od}.

As seen in section 4.2.5, a random utility model used to simulate path choice is given by:

$$p_{od,k} = p[k/od] = Prob[\, V_k - V_j \geq \varepsilon_j - \varepsilon_k \quad \forall j \in K_{od}] \quad \forall od,k$$
$$\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od}) \quad \forall od$$

where ε_j denotes the random residual corresponding to the perceived utility of path *j*. If the random residuals are equal to zero, $\varepsilon_j = 0$, i.e. the variance is null, then the variance-covariance matrix of the random residuals is null, $\Sigma = 0$, and the resulting choice model is deterministic. On the other hand, if it is assumed that the variance-

covariance matrix of the random residuals is non-null and non-singular, $|\boldsymbol{\Sigma}| \neq 0$, the model is probabilistic (see section 3.2).

Combining the path choice model with the systematic utility specification, a relation between choice probabilities and path costs for the *od* pair, known as the path choice map, is obtained:

$$p_{od,k} = p_{od,k}(\boldsymbol{V}_{od} = -\boldsymbol{g}_{od}) \;\forall od,k$$
$$\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od} = -\boldsymbol{g}_{od}) \quad \forall od$$

The above relation can be expressed using matrix notation (Fig. 5.2.4). Let

$\boldsymbol{P}$ be the path choice probabilities matrix, with a column for each *od* pair and a row for each path k the element k,od is given by $p[k/od]$ if path k connects the *od* pair, otherwise it is null ($\boldsymbol{P}$ is a block diagonal matrix with blocks given by the vectors $\boldsymbol{p}_{od}$).

Also, from the previous equations it follows that the matrix $\boldsymbol{P}$ depends on the path cost vector:

$$\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{V} = -\boldsymbol{g})$$

The flow h_k on the path k connecting the *od* pair, $k \in K_{od}$, is given by the product of the corresponding demand flow d_{od} and the path choice probability:

$$h_k = d_{od}\, p_{od,k}$$

and is measured in demand units. Thus, for each *od* pair, the relationship between path flows, path choice probabilities and demand flows is given by:

$$\boldsymbol{h}_{od} = d_{od}\, \boldsymbol{p}_{od}(\boldsymbol{V}_{od}) \quad \forall od \qquad (5.2.6a)$$

The relation (5.2.6a) for all O-D pairs can be expressed in aggregate form as:

$$\boldsymbol{h} = \boldsymbol{P}(\boldsymbol{V})\, \boldsymbol{d} \qquad (5.2.6b)$$

The whole demand model is defined by the relations (5.2.5-6), which combined together describe the relationship between path flows and path costs:

$$\boldsymbol{h}_{od} = d_{od}\, \boldsymbol{p}_{od}(-\boldsymbol{g}_{od}) \quad \forall od \qquad (5.2.7a)$$
$$\boldsymbol{h} = \boldsymbol{P}(-\boldsymbol{g})\, \boldsymbol{d} \qquad (5.2.7b)$$

For the usually-adopted probabilistic path choice models (with $|\Sigma| \neq 0$) (see Section 4.2.5), the demand model (5.2.7) is specified by a continuous function of path costs with continuous first derivatives. An example is reported in Fig. 5.2.4a.

In the case of a deterministic path choice model (with $\Sigma = \mathbf{0}$, see Section 3.2) a one-to-many map is usually obtained, since if there are several minimum-cost paths between an *od* pair the choice probabilities vector, $\boldsymbol{p}_{DET,od}$, and therefore the path flows vector, $\boldsymbol{h}_{DET,od}$, are not uniquely defined. An example is given in Fig. 5.2.4b.

$$g^T = [6 \quad 4 \quad 2 \quad 4 \quad 2 \quad 1]$$

$$p_{od,k} = \frac{\exp(-g_{od,k}/\theta)}{\sum_{j \in K_{od}} \exp(-g_{od,j}/\theta)};$$

$$\theta = 2$$

$$\boldsymbol{p}_{14} = \begin{bmatrix} 0.090 \\ 0.245 \\ 0.665 \end{bmatrix}; \quad \boldsymbol{p}_{24} = \begin{bmatrix} 0.269 \\ 0.731 \end{bmatrix}; \quad \boldsymbol{p}_{34} = [1.000]$$

P Matrix

Coppie O-D / Percor	1-4	2-4	3-4
1	0.090	0	0
2	0.245	0	0
3	0.665	0	0
4	0	0.269	0
5	0	0.731	0
6	0	0	1.000

$$\boldsymbol{h} = \boldsymbol{P} \cdot \boldsymbol{d}$$

$$\begin{bmatrix} 90 \\ 245 \\ 665 \\ 404 \\ 1097 \\ 800 \end{bmatrix} = \begin{bmatrix} 0.090 & 0 & 0 \\ 0.245 & 0 & 0 \\ 0.665 & 0 & 0 \\ 0 & 0.269 & 0 \\ 0 & 0.731 & 0 \\ 0 & 0 & 1.000 \end{bmatrix} \cdot \begin{bmatrix} 1000 \\ 1500 \\ 800 \end{bmatrix}$$

Fig. 5.2.4a – Example of demand model with probabilistic path choice.

$$\boldsymbol{g}^T = [6 \quad 4 \quad 2 \quad 4 \quad 2 \quad 1]$$

$$p_{od,k} \begin{cases} \in [0,1] & \text{se } g_{od,k} = \min_{j \in k_{od}} g_{od,j} \\ = 0 & \text{se } g_{od,k} > \min_{j \in k_{od}} g_{od,j} \end{cases}$$

$$\Sigma_{k \in k_{od}} p_{od,k} = 1$$

$$\boldsymbol{p}_{14} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}; \quad \boldsymbol{p}_{24} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}; \quad \boldsymbol{p}_{34} = [1]$$

P Matrix

Coppie O-D / Percorsi	1-4	2-4	3-4
1	0	0	0
2	0	0	0
3	1	0	0
4	0	0	0
5	0	1	0
6	0	0	1

$$\boldsymbol{h}_{DET} = \boldsymbol{P} \cdot \boldsymbol{d}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1000 \\ 0 \\ 1500 \\ 800 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1000 \\ 1500 \\ 800 \end{bmatrix}$$

Fig. 5.2.4b – Example of demand model with deterministic path choice

It can be useful to reformulate the deterministic demand model (5.2.7) through a system of inequalities. This system is obtained by applying to each *od* pair condition (3.5.11a) on the deterministic choice probabilities, $\boldsymbol{p}_{DET,od}$, described in Section 3.5 and repeated here for the convenience of the reader:

$$(\boldsymbol{V}_{od})^T(\boldsymbol{p}_{od} - \boldsymbol{p}_{DET,od}) \leq 0 \qquad \forall \boldsymbol{p}_{od} : \boldsymbol{p}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T\boldsymbol{p}_{od} = 1 \qquad \forall od$$

Since $\boldsymbol{V}_{od} = -\boldsymbol{g}_{od}$, multiplying this inequality by $d_{od} \geq 0\ \forall od$ yields:

$$\boldsymbol{g}_{od}^T(\boldsymbol{h}_{od} - \boldsymbol{h}_{DET,od}) \geq 0 \quad \forall \boldsymbol{h}_{od} : \boldsymbol{h}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T\boldsymbol{h}_{od} = d_{od} \qquad \forall od \qquad (5.2.7c)$$

Condition (5.2.7c) underlies the deterministic assignment models that will be described in Sections 5.3.2 and 5.4.2.

The deterministic demand model corresponds to a condition where "for each O-D pair the cost of the paths actually used is equal, and it is less or equal to the cost of each path not used":

$$h_{DET,k} > 0 \ \Rightarrow\ g_k = min\,(\boldsymbol{g}_{od}) \qquad k \in K_{od}$$

that is

$$g_k > min\,(\boldsymbol{g}_{od}) \ \Rightarrow\ h_{DET,k} = 0 \qquad k \in K_{od}$$

In the literature, the above condition is known as *Wardrop's first principle.*

The above inequalities are equivalent to the definition of the deterministic path choice model reported in Section 4.2.5; thus, the probability $p_{od,k}$ of a user of the *od* pair choosing path k is strictly positive only if the cost of path k is less than or equal to the cost of any other path that connects the *od* pair.

5.2.3. Feasible path and link flow sets

The vectors of non-negative path flows $\boldsymbol{h}$ compatible with the network topology and with the demand flows $\boldsymbol{d}$ are said to be *feasible.* The set S_h of feasible path flows contains non-negative vectors, $\boldsymbol{h} \geq \boldsymbol{0}$, such that for each *od* pair the sum of the elements of (sub-)vector $\boldsymbol{h}_{od}$ is equal to the demand flow:

$$\Sigma_{k \in Iod}\, h_{od,k} = d_{od}$$

or

$$\boldsymbol{1}^T\boldsymbol{h}_{od} = d_{od}$$

The set S_h of the vectors of feasible path flows therefore can be formally expressed as:

$$S_h = \{\boldsymbol{h} : \boldsymbol{h}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T\boldsymbol{h}_{od} = d_{od}\ \forall od\} \qquad (5.2.8)$$

The set S_h is compact since it is closed, and limited because the elements of the path flow vectors for each *od* pair belong to the interval $[0,d_{od}]$. It is also convex since it is defined by linear equations and inequalities. Furthermore, it is non-empty if at least one path for each *od* pair is available. Moreover, by definition, regardless of the path costs vector $\boldsymbol{g}$, the result of the demand model (5.2.7) is still a vector of feasible path flows:

$$\boldsymbol{h} = \boldsymbol{P}(\boldsymbol{-g})\boldsymbol{d} \in S_h \quad \forall \boldsymbol{g}$$

As for path flows, a non-negative link flow vector is feasible if it is compatible with the network topology and the demand flows $\boldsymbol{d}$. Thus, a vector of link flows $\boldsymbol{f}$ is feasible if, according to the supply model (5.2.3), it corresponds to feasible path flows as defined in the demand model. The set S_f of feasible link flows can be formally expressed[4] as:

$$S_f = \{\boldsymbol{f} : \boldsymbol{f} = \boldsymbol{\Delta h},\ \forall \boldsymbol{h} \in S_h\} \tag{5.2.9a}$$

that is

$$S_f(\boldsymbol{d}) = \{\boldsymbol{f} : \boldsymbol{f} = \Sigma_{od} \boldsymbol{\Delta}_{od} \boldsymbol{h}_{od},\ \boldsymbol{h}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T \boldsymbol{h}_{od} = d_{od}\ \forall od\} \tag{5.2.9b}$$

Formulation (5.2.9b) shows the role of the demand flows vector, $\boldsymbol{d}$, in the definition of the set of the feasible link flows S_f. The set S_f is (not empty) compact and convex since it is obtained through a linear transformation of the set of feasible path flow vectors, which has these characteristics (see Appendix A) assuming that the set of available paths for each *od* pair is finite.

It should be noted that, in general, there are more paths than links (that is the incidence matrix, $\boldsymbol{\Delta}$, has more columns than rows and is therefore non-invertible). Therefore, it is likely that several feasible path flow vectors lead to the same feasible link flow vector.

5.2.4. Network performance indicators

Each pattern of path and link costs and flows can be associated with indicators referring to an O-D pair or to the system as a whole, which will be used in the following sections. In particular, the *total cost* of the *od* pair, TC_{od}, is given by the sum of the products between the corresponding path costs and flows:

$$TC_{od} = \Sigma_{k \in Kod}\, h_k g_k = (\boldsymbol{g}_{od})^T \boldsymbol{h}_{od} \quad \forall od$$

to which corresponds an average cost, AC_{od}, obtained by dividing by the demand flow:

$$AC_{od} = TC_{od} / d_{od} = (\boldsymbol{g}_{od})^T \boldsymbol{h}_{od} / d_{od} \quad \forall od$$

The total cost for the whole network, TC, is given by the sum of the total costs over all the O-D pairs:

$$TC = \Sigma_{od} TC_{od} = \Sigma_{od} \Sigma_{k \in Kod} h_k g_k = \Sigma_k h_k g_k = \boldsymbol{g}^T \boldsymbol{h}$$

to which corresponds an average cost, AC, obtained by weighting the average costs of all the O-D pairs by the demand flows, that is by weighting the path costs by the path flows:

$$AC = (\Sigma_{od} AC_{od}\ d_{od}) / (\Sigma_{od} d_{od}) = (\Sigma_{od} \Sigma_{k \in Kod} h_k g_k) / (\Sigma_{od} \Sigma_{k \in Kod} h_k) =$$
$$= (\Sigma_{od} TC_{od}) / (\Sigma_{od} d_{od}) = TC / d_{..} = \boldsymbol{g}^T \boldsymbol{h} / \mathbf{1}^T \boldsymbol{h}$$

where $d_{..} = \Sigma_{od} d_{od} = \Sigma_{od} \Sigma_{k \in Kod} h_k = \mathbf{1}^T \boldsymbol{h}$ gives the total demand flow.

With reference to the additive and non-additive path costs, the following also holds:

$$TC = (\boldsymbol{g}^{ADD})^T \boldsymbol{h} + (\boldsymbol{g}^{NA})^T \boldsymbol{h} = (\Delta^T \boldsymbol{c})^T \boldsymbol{h} + (\boldsymbol{g}^{NA})^T \boldsymbol{h} = \boldsymbol{c}^T \boldsymbol{f} + (\boldsymbol{g}^{NA})^T \boldsymbol{h}$$

an expression, which in the case of null non-additive path costs $\boldsymbol{g}^{NA} = \mathbf{0}$, reduces to:

$$TC = \boldsymbol{c}^T \boldsymbol{f} = \Sigma_l f_l c_l \qquad (5.2.10)$$

In other words, the sum of the link costs multiplied by the corresponding flows coincides with the total network cost in the absence of non-additive costs.

To each *od* pair, dependent on the path choice model adopted, an *Expected Maximum Perceived Utility* (or *EMPU*), s_{od}, can be associated (See Section 3.5). This expected utility is a function of the systematic utility of the paths (without including other attributes, i.e. V°_{od}, for the sake of simplicity):

$$s_{od} = s_{od}(\boldsymbol{V}_{od} = -\boldsymbol{g}_{od}) = s_{od}(-\Delta_{od}^T \boldsymbol{c} - \boldsymbol{g}_{od}^{NA}) \quad \forall od \qquad (5.2.11)$$

Note that (see section 3.5) the EMPU is greater than or equal to the maximum systematic utility and therefore the average systematic utility:

$$s_{od} \geq max(V_{od}) \geq (\boldsymbol{V}_{od})^T \boldsymbol{p}_{od} = (\boldsymbol{V}_{od})^T \boldsymbol{h}_{od} / d_{od} \quad \forall od$$

The EMPU is therefore greater than or equal to the negative of the minimum cost over all the paths, which in turn is greater than or equal to the negative of the average cost:

$$s_{od} \geq -min(\boldsymbol{g}_{od}) \geq -(\boldsymbol{g}_{od})^T \boldsymbol{h}_{od} / d_{od} = -AC_{od} \quad \forall od$$

The total EMPU, TS, is defined as the sum of the EMPU of each O-D pair multiplied by the corresponding demand flow:

$$TS = \Sigma_{od}\, d_{od}\, s_{od}(V_{od}) = \Sigma_{od}\, d_{od}\, s_{od}(-\boldsymbol{g}_{od}) = \Sigma_{od}\, d_{od}\, s_{od}(-\Delta_{od}^{T}\boldsymbol{c} - \boldsymbol{g}_{od}^{NA})$$

The corresponding average EMPU, AS, is obtained by dividing by the total demand flow:

$$AS = \Sigma_{od}\, d_{od}\, s_{od} / \Sigma_{od}\, d_{od} = \Sigma_{od}\, d_{od}\, s_{od} / d_{..} = TS / d_{..}$$

In conclusion, the total cost is an estimate, carried out without considering the effect of dispersion, of the disutility users receive when distributing among paths according to path flows $\boldsymbol{h}$, while the EMPU is the disutility users perceive when making path choice leading to path flows $\boldsymbol{h}$ including the effect of dispersion.

From the preceding considerations, the following relations hold between total or average EMPU, and the opposite of the total or average cost respectively:

$$TS \geq -TC \qquad AS \geq -AC$$

Examples of the calculation of network indicators are reported in Fig. 5.2.5.

O-D Pair	Path	Cost	Flow	Total cost	Average Cost	-min (g)	exp(-C/θ)	Average EMPU $s=\theta\ln(\Sigma\exp(-C/\theta))$	Total EMPU
1-4	1	6	90	540			0.00248		
	2	4	245	980			0.01832		
	3	**2**	665	1330			0.13534		
	Total		1000	2850			0.15613		-1857
					2.85	2.00		-1.85	
2-4	4	4	404	1616			0.01832		
	5	**2**	1096	2192			0.13534		
	Total		1500	3808			0.15365		2810
					2.54	2.00		-1.87	
3-4	6	**1**	800	800			0.36788		
	Total		800	800			0.36788		800
					1.00	1.00		-1.00	
Total network values			3300	7458					5467
Average network values					2.26	1.75		-1.66	

Fig. 5.2.5 – Indicators for the network in Fig. 5.2.1.

As an example, the preceding relationships are applied to two different path choice models for which the EMPU can be calculated in closed form. In particular, a Logit path choice model with parameter θ_{od} gives (see Section 3.5):

$$TS = \Sigma_{od}\, d_{od}\ \theta_{od}\ ln(\Sigma_{k\in Kod}\, exp(V_k/\theta_{od})) = \Sigma_{od}\, d_{od}\ \theta_{od}\ ln(\Sigma_{k\in Kod}\, exp(-g_k/\theta_{od})) \geq$$
$$\geq -\Sigma_{od}\, d_{od}\ min(\boldsymbol{g}_{od}) \geq -\Sigma_{od}\, d_{od}\ \Sigma_{k\in Kod}\, g_k\, (h_k / d_{od}) = -TC$$

In the case of a deterministic path choice model, the EMPU is equal both to the maximum systematic utility and the average systematic utility (Section 3.5), thus the total EMPU is equal to the negative of total cost:

$$TS = \Sigma_{od}\, d_{od}\ max(\boldsymbol{V}_{od}) = \Sigma_{od}\, d_{od}\, \boldsymbol{V}_{od}{}^{T}\, \boldsymbol{p}_{od} = \Sigma_{od}\, d_{od}\ \boldsymbol{V}_{od}{}^{T}\, (\boldsymbol{h}_{od}/\, d_{od}) = -TC$$

since, in this case, the choice probability $\boldsymbol{p}_{od}$ vector and therefore the path flow vector $\boldsymbol{h}_{od}$ may have non-null elements only for minimum cost paths (Section 4.2.5).

5.3. Models for assignment to Uncongested Networks

Assignment to uncongested networks is based on the assumptions that flows and costs are mutually consistent and costs do not depend on flows[5]. In other words, path flows, and thus link flows, are obtained from path choice probabilities computed on the basis of flow-independent link performance measures and costs. In this sense, flows are consistent with costs and uncongested network assignment models follow a user equilibrium approach to demand-supply interaction. In the remainder of this chapter, however, the term equilibrium will be used only for congested network assignment, following common practice in the literature.

Uncongested assignment models are used for the analysis of relatively uncongested road transportation systems (generally, link cost functions are almost flat with respect to flows for flow-capacity ratios up to values of 0,50÷0,70). They are also used for analyzing public transport systems for which costs may be assumed independent of link passenger flows, if the available capacity is sufficient. Uncongested network assignment models, furthermore, are a component of congested network assignment models, which will be described in the following sections. Uncongested Network (*UN*) assignment models are defined by the demand model (5.2.7) expressing path flows as function of path costs and demand flows:

$$\boldsymbol{h}_{UN,od} = \boldsymbol{h}_{UN,od}\,(\boldsymbol{g}_{od};\, d_{od}) = d_{od}\boldsymbol{p}_{od}(-\boldsymbol{g}_{od})\ \ \forall \boldsymbol{g}_{od}\ \ \forall od \qquad (5.3.1a)$$
$$\boldsymbol{h}_{UN} = \boldsymbol{h}_{UN,}(\boldsymbol{g};\, \boldsymbol{d}) = \boldsymbol{P}(-\boldsymbol{g})\ \boldsymbol{d}\ \ \forall \boldsymbol{g} \qquad (5.3.1b)$$

It is possible to obtain path costs $\boldsymbol{g}$ from link costs $\boldsymbol{c}$ with equation (5.2.1), while the link flows $\boldsymbol{f}$ corresponding to the path flows $\boldsymbol{h}$ are given by equation (5.2.3). Fig. 5.3.1 depicts these relationships graphically, applying the scenario in Fig. 5.1.1 to the case of uncongested network assignment.

General uncongested network assignment models can be expressed in terms of link variables by combining equation (5.3.1) with (5.2.1) and (5.2.3). The result is called the uncongested network assignment map. This map associates a link flow vector to each demand flow vector and link cost vector, and can be expressed in an aggregate or disaggregate way as:

$$\boldsymbol{f}_{NL} = \boldsymbol{f}_{NL}(\boldsymbol{c}; \boldsymbol{d}) = \Sigma_{od}\, d_{od}\, \boldsymbol{\Delta}_{od}\, \boldsymbol{p}_{od}(-\boldsymbol{\Delta}_{od}^{T}\, \boldsymbol{c} - \boldsymbol{g}_{od}^{NA}) \quad \forall \boldsymbol{c} \qquad (5.3.2a)$$

$$\boldsymbol{f}_{NL} = \boldsymbol{f}_{NL}(\boldsymbol{c}; \boldsymbol{d}) = \boldsymbol{\Delta P}(-\boldsymbol{\Delta}^{T}\, \boldsymbol{c} - \boldsymbol{g}^{NA})\, \boldsymbol{d} \quad \forall \boldsymbol{c} \qquad (5.3.2b)$$

Note that link flows depend non-linearly on the link costs, but linearly on the demand flows, so that the effect of each O-D pair can be evaluated separately.

In the following sections, probabilistic and deterministic path choice models, which lead respectively to stochastic and deterministic uncongested network assignment models, will be described separately.

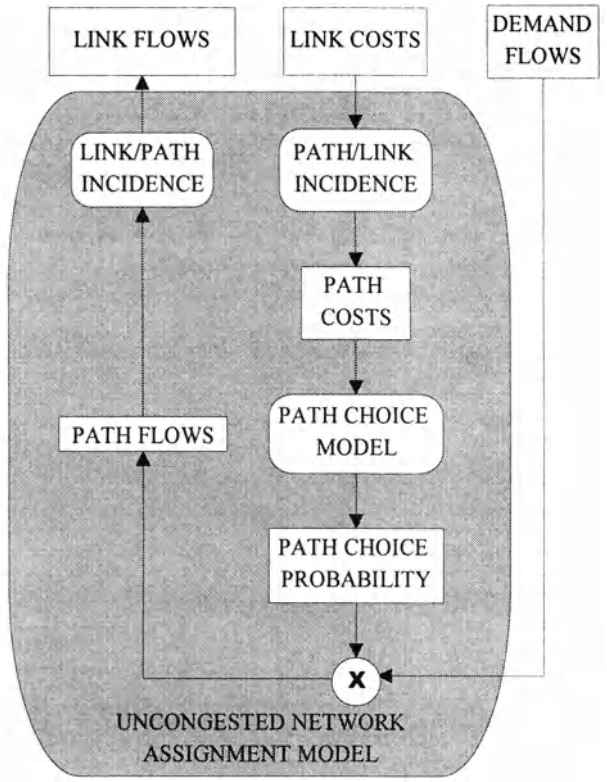


Fig. 5.3.1 – Schematic representation of uncongested network assignment models.

5.3.1. Models for Stochastic Uncongested Network assignment

If path choice behavior is simulated through a probabilistic random utility model, the resulting assignment model is known as a Stochastic Uncongested Network (*SUN*)[6] assignment. In this case, the resulting link or path flows correspond to a situation in which "for each O-D pair the *perceived* cost of the used paths is less than or equal to the cost of every other path". Using the probabilistic path choice models studied in Section 4.2.5, recall that each vector of link and path costs corresponds to a unique choice probability vector. It follows that the uncongested assignment map, usually defined by equations (5.3.2), is given by the *stochastic uncongested assignment function*, $f_{SUN}(c;d)$, a one-to-one correspondence associating any vector of link costs $\boldsymbol{c}$ with a vector of link flows $\boldsymbol{f}$ belonging to the non-empty, compact and convex set of feasible link flows (Fig. 5.3.2):

$$\boldsymbol{f}_{SUN} = \boldsymbol{f}_{SUN}(\boldsymbol{c};\boldsymbol{d}) = \Sigma_{od}\, d_{od}\, \Delta_{od}\, \boldsymbol{p}_{od}(-\Delta_{od}^{T}\boldsymbol{c} - \boldsymbol{g}_{od}^{NA}) \in S_f \;\forall \boldsymbol{c} \qquad (5.3.3)$$

Formulations of SUN analogous to (5.3.2b) and (5.3.1a,b) in terms of path costs and flows are possible, but will not be presented here for the sake of brevity.

The parameters of the stochastic uncongested assignment function, excluding the demand vector, are those of the path choice model (such as coefficients of the systematic utility and variance of the random residuals) and those of the supply model (such as travel times and generalized costs, as well as the graph topology). Under certain assumptions on the path choice function, the function (5.3.3) shows some features which will be useful in the analysis of stochastic equilibrium assignment models, and for this reason will be described in Section 5.4.1.

Variance and covariance of link and path flows, considered as random variables. Assuming probabilistic choice behavior (with known demand flows d_{od}) and independent user path choices, path flows can be considered realizations of multinomial random variables $\boldsymbol{H}_{od}$. The values, $\boldsymbol{h}_{od}$, calculated with the stochastic uncongested network assignment model, represent the means of $\boldsymbol{H}_{od}$, as shown at the beginning of Section 4.5 in the most general case of demand models on all choice dimensions. Therefore, the mean, variance and covariance of the elements of the path flow random vector, $\boldsymbol{H}$, can be expressed as:

$$\begin{aligned}
E[H_k] &= h_{SUN,k} = d_{od}\, p_{od,k} && \forall od,k \\
Var[H_k] &= d_{od}\, p_{od,k}\,(1 - p_{od,k}) && \forall od,k \\
Cov[H_k, H_j] &= \begin{cases} -\, d_{od}\, p_{od,k}\, p_{od,j} & k, j \in K_{od} \\ 0 & \text{otherwise} \end{cases} && \forall od,k,j
\end{aligned}$$

The first equation expresses the elements of the mean vector, $h_{SUN} = E[\boldsymbol{H}]$, of the random vector $\boldsymbol{H}$, while the last two equations give the elements of the variance-covariance matrix, Σ_H. If the path flows vector, $\boldsymbol{h}$, is considered the realization of a random vector, $\boldsymbol{H}$, the link flows vector, $\boldsymbol{f} = \Delta\boldsymbol{h}$, obtained from this with a linear

transformation, is a realization of a random vector, $\boldsymbol{F}$. Thus the mean vector, $E[\boldsymbol{F}] = \Delta\, E[\boldsymbol{H}] = \Delta \boldsymbol{h}_{SUN} = \boldsymbol{f}_{SUN}$, and the variance-covariance matrix, $\Sigma_F = \Delta^T \Sigma_H \Delta$, of the link flows random vector, $\boldsymbol{F}$, can be expressed in terms of the corresponding values of path flows, $\boldsymbol{h}_{SUN} = E[\boldsymbol{H}]$ and Σ_H.

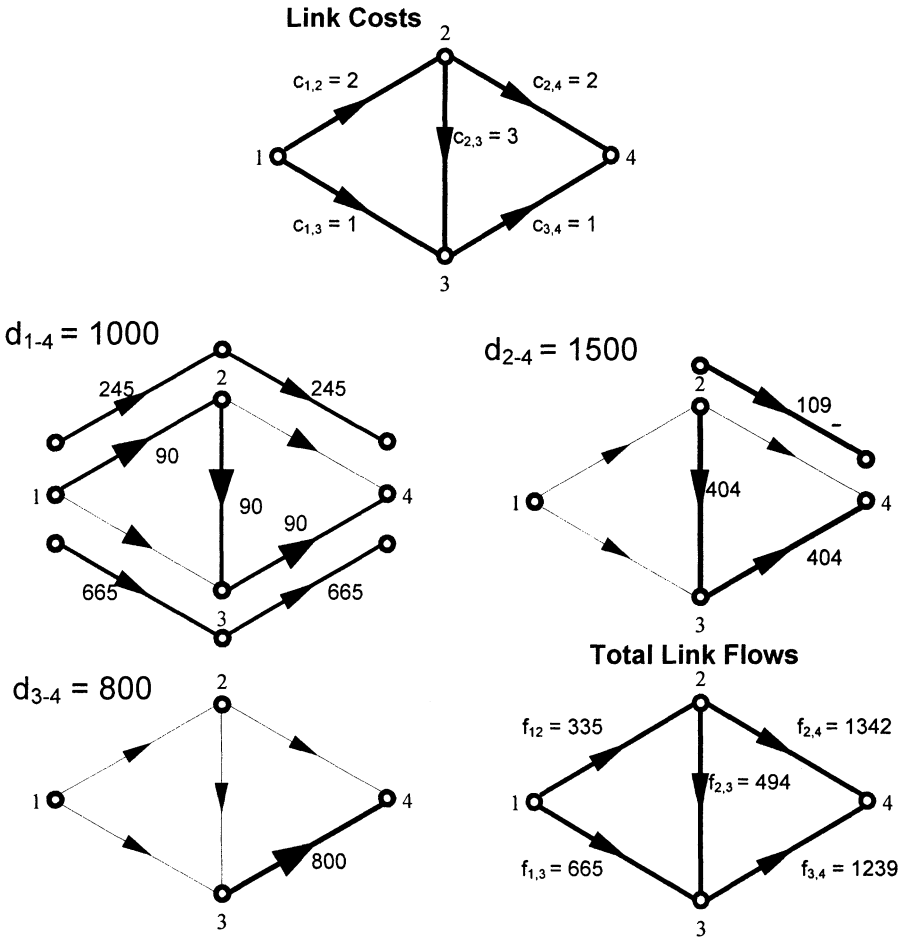


Fig. 5.3.2 - Stochastic Uncongested Network (*SUN*) assignment with the path choice model of Fig. 5.2.4a.

The link flow vector defined by the stochastic uncongested assignment function for a given link cost vector can easily be calculated in the case of path choice models based on explicit path enumeration. Alternatively, there are algorithms described in Chapter 7 which allow, for some path choice models, the efficient computation of link flows without explicit enumeration of paths.

5.3.2. Models for Deterministic Uncongested Network assignment

Under the assumption of deterministic path choice behavior, the demand flow of each O-D pair is assigned to the minimum cost path or paths (i.e. paths with maximum systematic utility), while no flow is assigned to other paths. Therefore, Deterministic Uncongested Network (DUN) assignment is also known as All-or-Nothing assignment[7]. In general, as has already been noted, several choice probability vectors may correspond to a single vector of link and path costs. It follows that the relationship (5.3.2), expressing the general uncongested network assignment, is detailed into the *deterministic uncongested network assignment map* $\boldsymbol{h}_{DUN} = \boldsymbol{h}_{DUN}(\boldsymbol{g}; \boldsymbol{d}) \in S_h$, which is a one-to-many (or point-to-set) map between path costs and flows. In other words, since there may be several alternative minimum cost paths connecting an origin to a destination, the same path and link costs vector may correspond to several vectors of deterministic uncongested networks path and link flows. For this reason, the study of the properties of deterministic network loading is preferably conducted using indirect formulations, equivalent to (5.3.2), based on the specification of the deterministic demand model with a system of inequalities (5.2.7c). Summing the inequalities (5.2.7c) over all *od* pairs yields expression (5.3.4):

$$\boldsymbol{g}^T(\boldsymbol{h} - \boldsymbol{h}_{DUN}) \geq 0 \quad \forall \boldsymbol{h} \in S_h \tag{5.3.4}$$

The resultant path (or link) flows correspond to the condition expressed by the *Wardrop principle* which states "for each O-D pair, the path cost used is equal, and is less than or equal to the cost of each unused path", as described in section 5.2.2. Figure 5.3.3 presents an example of the deterministic uncongested network assignment model.

If non-additive path costs are zero, $\boldsymbol{g}^{NA} = \boldsymbol{0}$, total path costs coincide with additive costs $\boldsymbol{g}^T = (\boldsymbol{g}^{ADD})^T = \boldsymbol{c}^T\boldsymbol{\Delta}$, and it is easy to verify that (5.3.4) is equivalent to:

$$\boldsymbol{c}^T(\boldsymbol{f} - \boldsymbol{f}_{DUN}) \geq 0 \qquad \forall \boldsymbol{f} \in S_f \tag{5.3.5}$$

On the other hand when there are non-additive path costs, expression (5.3.4) is equivalent to:

$$\boldsymbol{c}^T(\boldsymbol{f} - \boldsymbol{f}_{DUN}) + (\boldsymbol{g}^{NA})^T(\boldsymbol{g} - \boldsymbol{g}_{DUN}) \geq 0 \quad \forall \boldsymbol{f} = \boldsymbol{\Delta h}, \forall \boldsymbol{h} \in S_h \tag{5.3.6a}$$

In order to facilitate the analysis and solution (see Section 7.3.1) of model (5.3.6a), it can be reformulated without any explicit reference to path flows. Let:

$G^{NA} = (\boldsymbol{g}^{NA})^T\boldsymbol{h}$ be the total non-additive cost corresponding to the generic feasible vector of path flows $\boldsymbol{h}$;

$G^{NA}{}_{DUN} = (\boldsymbol{g}^{NA})^T\boldsymbol{h}_{DUN}$ be the total non-additive cost corresponding to the deterministic uncongested assignment of the path flow vector $\boldsymbol{h}_{DUN}$.

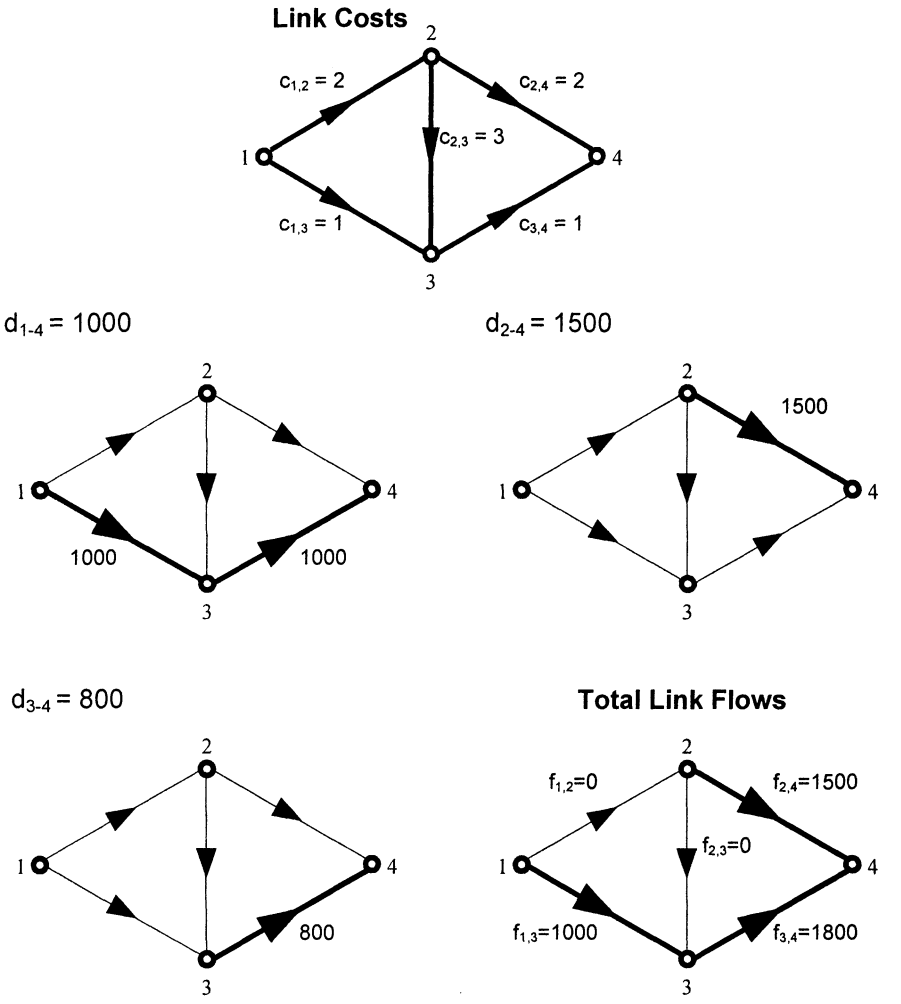


Fig. 5.3.3 Deterministic Uncongested Networks (DUN) assignment, with the path choice model in Fig. 5.2.4b.

In terms of link flows, $\boldsymbol{f}_{DUN}$, and of total non-additive cost, $G^{NA}{}_{DUN}$, corresponding to the deterministic uncongested network assignment[8], the following holds:

$$\boldsymbol{c}^T(\boldsymbol{f} - \boldsymbol{f}_{DUN}) + 1(G^{NA} - G^{NA}{}_{DUN}) \geq 0 \quad \forall \boldsymbol{f} = \boldsymbol{\Delta h},\ \forall G^{NA} = (\boldsymbol{g}^{NA})^T \boldsymbol{h} \quad \forall \boldsymbol{h} \in S_h \quad (5.3.6b)$$

The existence of solutions of the inequality systems (5.3.4,5,6) is assured since they are defined over limited feasible sets. Clearly, demand flows affect the solution

since they appear in the definition of the feasible sets over which the problems are defined.

Formulation with optimization models. Deterministic Uncongested Network assignment can also be formulated with an optimization model, or more precisely with a linear programming model. In fact, if the non-additive path costs are zero, it is easy to verify that the inequality system (5.3.5) is equivalent to an optimization model with linear objective function and a set of linear equality or inequality constraints as given below:

$$\begin{aligned} \boldsymbol{f}_{DUN}(\boldsymbol{c};\boldsymbol{d}) = \; & argmin \;\; \boldsymbol{c}^T\boldsymbol{f} \\ & \boldsymbol{f} \in S_f(\boldsymbol{d}) \end{aligned} \tag{5.3.7}$$

where the notation $S_f(\boldsymbol{d})$ highlights the role of the demand flow vector in the definition of the feasible link flow set. If there are non-additive path costs, the relation (5.3.7) becomes:

$$\begin{aligned} \boldsymbol{f}_{DUN}(\boldsymbol{c};\boldsymbol{d}),\, G^{NA}{}_{DUN} = \; & argmin \;\; \boldsymbol{c}^T\boldsymbol{f} + 1 \cdot G^{NA} \\ & \boldsymbol{f} = \Delta\boldsymbol{h} \\ & G^{NA} = (\boldsymbol{g}^{NA})^T\boldsymbol{h} \\ & \boldsymbol{h} \in S_h \end{aligned} \tag{5.3.8}$$

These formulations are most easily understood by considering that the assignment of each demand flow to a minimum cost path corresponds to the case where simultaneously "the cost for each user is minimum" and "the total network cost is minimum" (the link costs being independent of flows).

Regardless of the model adopted, the link flow vector (or rather one of the vectors) resulting from deterministic uncongested network assignment can easily be calculated when using path choice models based on explicit path enumeration. Without explicit path enumeration and when non-additive path costs are equal to zero, a link flow vector can easily be obtained with procedures based on algorithms for the calculation of minimum cost paths (see Chapter 7), or by directly solving optimization models (5.3.7-8)[9].

5.4. Rigid demand Users Equilibrium assignment models

In the case of congested networks, link performance measures and costs depend on link flows, through the performance and cost functions introduced in Chapter 2. On the other hand, link flows depend on link costs through path choice probabilities, as described by the uncongested network assignment map. The user equilibrium approach to the study of the supply-demand interactions assumes that a configuration of path flows $\boldsymbol{h}^*$ mutually consistent with the corresponding path costs $\boldsymbol{g}^*$ is representative of the state assumed by the real-world system[10]. Equilibrium path flows and costs are defined by a system of non-linear equations obtained by combining the supply model (5.2.4) with the demand model (5.2.7):

$$\boldsymbol{g}^* = \boldsymbol{\Delta}^T \boldsymbol{c}(\boldsymbol{\Delta h}^*) + \boldsymbol{g}^{NA}$$
$$\boldsymbol{h}^* = \boldsymbol{P}(-\boldsymbol{g}^*)\, \boldsymbol{d}$$

Equivalent equilibrium assignment models defined with link variables can be expressed by the system of non-linear equations obtained by combining the uncongested network assignment map (5.3.2) with the flow-dependent cost functions (5.2.2):

$$\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$$
$$\boldsymbol{f}^* = \boldsymbol{f}_{NL}(\boldsymbol{c}^*; \boldsymbol{d}) = \boldsymbol{\Delta P}(-\boldsymbol{\Delta}^T \boldsymbol{c}^* - \boldsymbol{g}^{NA})\, \boldsymbol{d}$$

The above system of equations shows that, in the case of congested networks, link flows may depend non-linearly on demand flows (unlike in uncongested network assignment). Thus, in this case, the effect of each O-D pair cannot be evaluated separately.

The circular dependence between flows and costs expressed by the equilibrium approach is depicted in Fig. 5.4.1. This figure shows the scenario in Fig. 5.1.1 for the case of rigid demand dealt with in this section, highlighting the role of the uncongested networks assignment model.

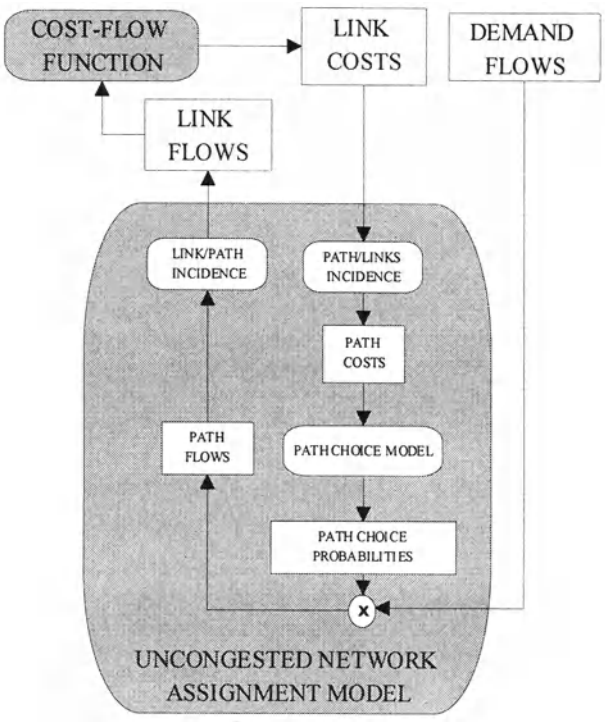


Fig. 5.4.1 – Schematic representation of rigid demand equilibrium assignment models.

The actual formulation and the analysis of the theoretical properties (existence and uniqueness) of equilibrium flows (and costs) depend on the type of model adopted to simulate path choices, probabilistic or deterministic. This selection defines respectively stochastic and deterministic equilibrium assignment models, separately described in the following sections.

5.4.1. Stochastic User Equilibrium models

Stochastic user equilibrium assignment (in the literature, SUE) is obtained by applying the equilibrium approach to congested networks under the assumption of probabilistic path choice behavior. The resulting path flows $\boldsymbol{h}^*$ correspond to the condition in which "for each O-D pair the perceived cost of paths used at the equilibrium is less than or equal to the perceived cost of every other path". Equilibrium path flows can be expressed as the solution of a fixed-point model defined on the feasible path flows set S_h, obtained by combining the supply model (5.2.4) with the demand model (5.2.7):

$$\boldsymbol{h}^* = \boldsymbol{P}(-\Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^*) - \boldsymbol{g}^{NA})\, \boldsymbol{d} \qquad (5.4.1)$$

with $\boldsymbol{h}^* \in S_h$

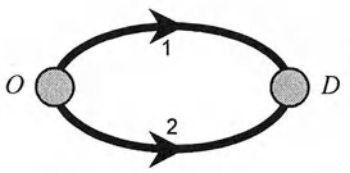
An equivalent fixed-point model using link flow variables $\boldsymbol{f}^*$ and therefore defined on the feasible link flows set S_f can be obtained by combining the stochastic uncongested network assignment function (5.3.3) (expressed in disaggregate form to facilitate the analysis) with the flow-dependent cost functions (5.2.2):

$$\begin{aligned} &\boldsymbol{f}^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{f}^*)) \quad \text{or} \\ &\boldsymbol{f}^* = \textstyle\sum_{od} d_{od}\, \Delta_{od}\, \boldsymbol{p}_{od}(-\Delta_{od}^T\, \boldsymbol{c}(\boldsymbol{f}^*) - \boldsymbol{g}_{od}^{NA}) \end{aligned} \qquad (5.4.2)$$

with $\boldsymbol{f}^* \in S_f$

The corresponding equilibrium costs can be obtained with the equations reported in section 5.2. Fixed-point models expressed in terms of link or path cost variables are also possible to develop.

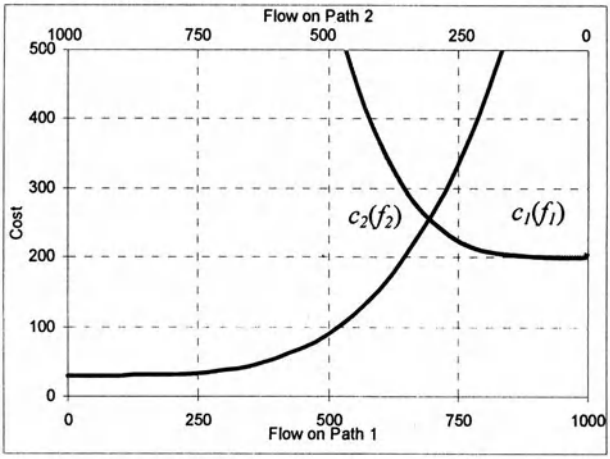
An example of stochastic equilibrium using a Logit path choice model for a two-link/path network is given in Fig. 5.4.2. The stochastic equilibrium pattern is obtained at the intersection of the curves representing the supply and (inverse) demand equations. Note that the stochastic equilibrium configuration does not correspond to equal (systematic) costs on the two paths, which means that the intersection point of the two curves does not correspond to the null value of the difference $g_1 - g_2$. In other words, at the stochastic equilibrium, some travelers that use higher (systematic) cost paths. This result obviously depends on the assumptions made on path choice behavior. The perceived path cost is modeled as a random variable and therefore some users may choose higher (systematic) cost paths because they perceive them as least cost.

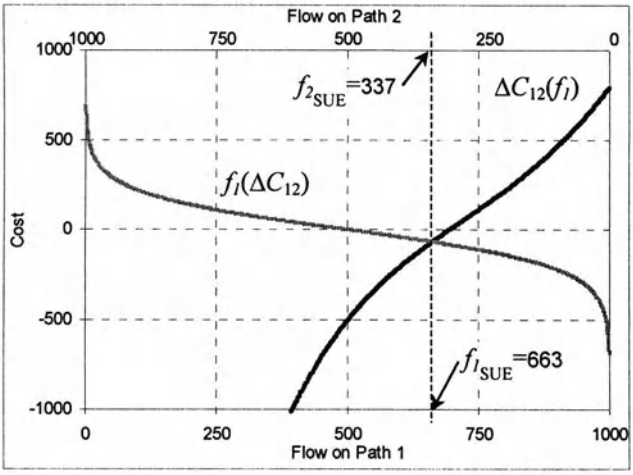


$$c_1 = c_{0,1}\left(1 + a_1 \frac{f_1}{Cap_1}\right)^{\gamma} = 30\left(1 + 2\left(\frac{f_1}{500}\right)^4\right);$$

$$c_2 = c_{0,2}\left(1 + a_2 \frac{f_2}{Cap_2}\right)^{\gamma} = 200\left(1 + 2\left(\frac{f_2}{500}\right)^4\right);$$

$$F_1 + F_2 = d_{od} = 1000; \quad F_1 = f_1; \quad F_2 = f_2;$$





Supply equation $\Delta C_{1,2}(f_1) = C_1(f_1) - C_2(f_2=d-f_1)$

Demand equation $f_1(\Delta C_{1,2}) = d_{o,d} \dfrac{1}{1 + exp(\Delta C_{1,2}/\theta)}$

$f_2 = d_{od} - f_1$

Fig. 5.4.2 – Example of Stochastic User Equilibrium (SUE). (θ= 100)

The existence and uniqueness of stochastic equilibrium flows and costs are guaranteed respectively by the continuity and the monotonicity of the cost functions under the rather general assumption that the path choice model guarantees the continuity and monotonicity of the stochastic uncongested network assignment function (as described below). Note that the conditions for existence and uniqueness described are only sufficient; that is, there can be non-continuous and/or non-monotone cost functions giving rise to a unique equilibrium configuration. In the following, existence and uniqueness are explicitly analyzed only for equilibrium link flows; these conditions ensure the existence and uniqueness of the corresponding link costs, $\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$, and of the path costs and flows, $\boldsymbol{g}^*$ and $\boldsymbol{h}^*$ obtained through relations (5.2.1) and (5.2.7) respectively.

Continuity of the stochastic uncongested network assignment function. If the path choice model is defined by a continuous function (with continuous first partial derivatives), as is the case for the probabilistic models usually adopted (with $|\Sigma| \neq 0$), the stochastic uncongested network assignment function is continuous (with continuous first partial derivatives) with respect to link costs. In other words, a "small" variation in link costs induces a "small" variation in link flows.

Existence of Stochastic User Equilibrium link flows. The fixed-point model (5.4.2) has at least one solution if the function of the path choice probabilities, $\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od})$ (which defines the stochastic uncongested network assignment function $\boldsymbol{f} = \boldsymbol{f}_{SUN}(\boldsymbol{c}; \boldsymbol{d})$) and the cost function $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are continuous.

In fact, the equilibrium solution $\boldsymbol{f}^*$ is a fixed point of the compound function $\boldsymbol{y} = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{x}))$, which under the above assumptions is a continuous function defined over the non-empty, compact, and convex set S_f (for a connected network) Furthermore, the function $\boldsymbol{y} = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{x}))$ assumes values only in the definition set S_f; thus all of the assumptions of Brouwer's theorem on the existence of fixed points are satisfied (see Appendix A).

The continuity of the cost functions over the feasible flow set (and therefore the existence of the equilibrium solution) requires that the cost functions are defined for any feasible value of link flows, even if a particular link flow is greater than the physical capacity of that link (links flows are bounded above by the demand flows). In fact, if explicit capacity constraints are added, the set of feasible flow might be empty. In other words, there may be no link flow vector that corresponds to the transportation demand and simultaneously does not exceed the capacity of each network link. Such a limit case corresponds to an excess of demand with respect to the available capacity of the system.

Monotonicity of the stochastic uncongested network assignment function. If the path choice model is defined by a non-decreasing monotone function of systematic utility, as in the case of additive probabilistic models (as demonstrated in Section

3.5), the stochastic uncongested network assignment function is non-increasing monotone with respect to the link costs. Thus, if the cost of one or more of the links increases, the flow (or flows) on these links decreases, and vice versa. This property is formally expressed as:

$$(\boldsymbol{f}_{SUN}(\boldsymbol{c}')-\boldsymbol{f}_{SUN}(\boldsymbol{c}''))^T(\boldsymbol{c}'-\boldsymbol{c}'')\le 0 \quad \forall \boldsymbol{c}', \boldsymbol{c}''$$

Given any two link cost vectors $\boldsymbol{c}'$ and $\boldsymbol{c}''$, consider the following notation:

$$\begin{aligned}
&\boldsymbol{g}_{od}' = \Delta_{od}^T\boldsymbol{c}' + \boldsymbol{g}_{od}^{NA} \qquad V_{od}' = -\boldsymbol{g}_{od}'\,\boldsymbol{p}_{od}' = \boldsymbol{p}_{od}(V_{od}')\\
&\boldsymbol{h}_{od}' = d_{od}\boldsymbol{p}_{od}'\boldsymbol{f}' = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od}'\\
&\boldsymbol{g}_{od}'' = \Delta_{od}^T\boldsymbol{c}'' + \boldsymbol{g}_{od}^{NA} \qquad V_{od}'' = -\boldsymbol{g}_{od}'' \qquad \boldsymbol{p}_{od}'' = \boldsymbol{p}_{od}(V_{od}'')\\
&\boldsymbol{h}_{od}'' = d_{od}\boldsymbol{p}_{od}''\,\boldsymbol{f}'' = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od}''
\end{aligned}$$

Assuming that the path choice model is additive and the choice map is non-decreasing monotone (see Section 3.5) yields:

$$(\boldsymbol{p}_{od}(V_{od}') - \boldsymbol{p}_{od}V_{od}'')^T(V_{od}' - V_{od}'') \ge 0 \;\forall od$$

and it follows from non-negativity of the demand flow $d_{od} \ge 0$ that:

$$\begin{aligned}
&d_{od}(\boldsymbol{p}_{od}(V_{od}') - \boldsymbol{p}_{od}V_{od}'')^T(V_{od}' - V_{od}'') \ge 0 \quad \forall od\\
&(\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T(V_{od}' - V_{od}'') \ge 0 \;\forall od\\
&\Sigma_{od}(\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T(V_{od}' - V_{od}'') \ge 0
\end{aligned}$$

Since $V_{od} = -\boldsymbol{g}_{od} = -\Delta_{od}^T\boldsymbol{c} - \boldsymbol{g}_{od}^{NA}$, the above reduces to:

$$\begin{aligned}
&-\Sigma_{od}(\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T(\boldsymbol{g}_{od}' - \boldsymbol{g}_{od}'') \ge 0\\
&\Sigma_{od}(\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T(\Delta_{od}^T\boldsymbol{c}' + \boldsymbol{g}_{od}^{NA} - \Delta_{od}^T\boldsymbol{c}'' - \boldsymbol{g}_{od}^{NA}) \le 0\\
&\Sigma_{od}(\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T\Delta_{od}^T(\boldsymbol{c}' - \boldsymbol{c}'') \le 0
\end{aligned}$$

from which $(\boldsymbol{f}' - \boldsymbol{f}'')^T(\boldsymbol{c}' - \boldsymbol{c}'') \le 0$ follows.

Note that two different vectors of link costs, $\boldsymbol{c}'$ and $\boldsymbol{c}''$ usually generate two different vectors of the additive path costs $\Delta_{od}{}^T\boldsymbol{c}'$ and $\Delta_{od}{}^T\boldsymbol{c}''$, and therefore two vectors of systematic utility, V_{od}' and V_{od}''. Thus, the assumption that the path choice model is additive (see Section 3.5) with respect to the path systematic utility is equivalent to assuming that for each od pair, the parameters of the distribution of the path random residuals ε_{od}, (such as the parameter θ in a Logit model or the variance-covariance matrix Σ in a Probit model) do not depend on the additive path costs, and therefore on the link costs relevant to congestion. However, they may

depend on other reference variables (such as distance, null flow costs, etc.)[11]. Note that under this assumption, the Jacobian of the function $\boldsymbol{f}_{SUN} = \boldsymbol{f}_{SUN}(\boldsymbol{c})$, $\boldsymbol{Jac}[\boldsymbol{f}_{SUN}(\boldsymbol{c})] = \sum_{od} d_{od}\, \Delta_{od}\, \boldsymbol{Jac}[\boldsymbol{p}_{od}(-\Delta_{od}^{T}\boldsymbol{c}-\boldsymbol{g}_{od}^{NA})]\Delta_{od}^{T}$ is symmetric and negative semi-definite, since the Jacobian $\boldsymbol{Jac}[\boldsymbol{p}_{od}(-\Delta_{od}^{T}\boldsymbol{c}-\boldsymbol{g}_{od}^{NA})]$ is symmetric positive semi-definite (see Section 3.5)

Uniqueness of Stochastic User Equilibrium link flows. The fixed-point model (5.4.2) has at most one solution if the link cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are strictly increasing[12] over the set of feasible link flows:

$$[\boldsymbol{c}(\boldsymbol{f}') - \boldsymbol{c}(\boldsymbol{f}'')]^{T}(\boldsymbol{f}' - \boldsymbol{f}'') > 0 \qquad \forall \boldsymbol{f}' \neq \boldsymbol{f}'' \in S_f$$

and the path choice models are additive (and expressed by continuous functions $\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od})$ with continuous first partial derivatives).

As previously shown, under this assumption the stochastic uncongested network assignment function $\boldsymbol{f}_{SUN}(\boldsymbol{c})$ is monotone non-increasing with the link costs:

$$[\boldsymbol{f}_{SUN}(\boldsymbol{c}') - \boldsymbol{f}_{SUN}(\boldsymbol{c}'')]^{T}(\boldsymbol{c}' - \boldsymbol{c}'') \leq 0 \quad \forall \boldsymbol{c}', \boldsymbol{c}''$$

The proof is then completed by *reductio ad absurdum*. If two different equilibrium link flow vectors existed, $\boldsymbol{f}_1^* \neq \boldsymbol{f}_2^* \in S_f$, assuming $\boldsymbol{c}_1^* = \boldsymbol{c}(\boldsymbol{f}_1^*)$ and $\boldsymbol{c}_2^* = \boldsymbol{c}(\boldsymbol{f}_2^*)$, the equilibrium definition $\boldsymbol{f}_1^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}_1^*)$ and $\boldsymbol{f}_2^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}_2^*)$ and the monotonicity of the stochastic uncongested network assignment function with $\boldsymbol{c}' = \boldsymbol{c}_1^*$ and $\boldsymbol{c}'' = \boldsymbol{c}_2^*$ yields:

$$[\boldsymbol{f}_1^* - \boldsymbol{f}_2^*]^{T}(\boldsymbol{c}_1^* - \boldsymbol{c}_2^*) \leq 0$$

From the monotonicity of the cost functions, with $\boldsymbol{f}' = \boldsymbol{f}_1^* \neq \boldsymbol{f}'' = \boldsymbol{f}_2^*$, it follows that:

$$[\boldsymbol{c}_1^* - \boldsymbol{c}_2^*]^{T}(\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) > 0$$

Thus, there is a contradiction between the monotonicity of the cost functions and that of the stochastic uncongested network assignment function.

A sufficient condition for the strict monotonicity of the cost functions is that the Jacobian matrix $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$ of the cost vector $\boldsymbol{c}(\boldsymbol{f})$ is positive definite over the set S_f (see Appendix A). In the case of separable cost functions, $c_l = c_l(f_l)$, the Jacobian matrix is diagonal and the its elements are the derivatives of the cost functions of each link with respect to link flow. In the usual case when cost functions are increasing with respect to the flow[13], the derivatives are positive, the Jacobian

matrix is positive definite, and the equilibrium flow vector $\boldsymbol{f}^*$ is unique. However, there are real situations in which the cost functions are not monotone.

In applications, the non-uniqueness of equilibrium, or the difficulty of demonstrating it *a priori*, gives rise to problems in both computation and interpretation. On one hand, it is possible to demonstrate the convergence of the solution algorithms only if the solution is unique (see Section 7.4). On the other hand, the inability of demonstrating the uniqueness of equilibrium implies that the equilibrium flow vector that has been calculated is not necessarily the one with which to design/evaluate the transportation system under study. In other words, in the context of the functioning of the system, in the latter case the system may attain different equilibrium patterns, and each of these should be verified.

Stochastic equilibrium link flows can be calculated with various algorithms, the simplest of which use the stochastic uncongested network assignment function as described in Chapter 7. Appendix 5A at the end of this chapter proposes some formulations of SUE with rigid demand based on optimization models. These models can be used to specify other algorithms derived from general optimization techniques.

5.4.2. Deterministic User Equilibrium models

Deterministic user equilibrium assignment is obtained by applying the equilibrium approach for congested networks under the assumption of deterministic path choice behavior. In the literature, this problem is usually denoted by the acronym *DUE* (Deterministic User Equilibrium). Deterministic equilibrium link flows, $\boldsymbol{f}^*$, path flows, $\boldsymbol{h}^*$, and the corresponding costs, $\boldsymbol{c}^*$ or $\boldsymbol{g}^*$, can be determined with a fixed-point model obtained by simultaneously applying the supply model (5.2.4) and the demand model (5.2.7), as in the stochastic equilibrium case (an alternative is to utilize the deterministic uncongested network assignment map and flow-dependent cost functions). In this case, however, there are some mathematical complications arising from the fact that the deterministic demand model is expressed (like the corresponding deterministic uncongested network assignment map[14]) by a one-to-many map, as was noted in Section 5.2.2 (and in 5.3.2).

For this reason, the properties of deterministic equilibrium are usually studied through indirect formulations. The most general is the *variational inequality* formulation based on the specification of the deterministic demand model with the system of inequalities (5.2.7c):

$$\boldsymbol{g}(\boldsymbol{h}^*)^T(\boldsymbol{h} - \boldsymbol{h}^*) \geq 0 \quad \forall \boldsymbol{h} \in S_h \tag{5.4.3}$$

By combining the demand model obtained by summing (5.2.7c) on all the *od* pairs with the supply model (5.2.4), expression (5.4.3) is obtained.

In the case of congested networks, therefore, the resultant path (or link) flows correspond to the condition expressed by the *Wardrop first principle*: "for each O-D

pair the path *equilibrium* cost used is equal, and is less than or equal to the *equilibrium* cost of each unused path" (see Section 5.2.2).

Equivalent variational inequality models expressed in terms of link flows are based on combining the link cost functions (5.2.2) with the inequality systems (5.3.5) or (5.3.6) representing the deterministic uncongested network assignment:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f}-\boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f \tag{5.4.4}$$

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f}-\boldsymbol{f}^*) + (\boldsymbol{g}^{NA})^T(\boldsymbol{h}-\boldsymbol{h}^*) \geq 0 \quad \forall \boldsymbol{f}=\Delta\boldsymbol{h} \ \forall \boldsymbol{h} \in S_h \tag{5.4.5a}$$

The expressions (5.4.4) and (5.4.5a) apply respectively to cases with null and non-null non-additive path costs. Note that the expressions (5.4.3-5) are different from those used for deterministic uncongested assignment in that path and link costs depend on flows. In the presence of non-additive path costs, the considerations made in Section 5.3.2 hold and (5.4.5a) can be expressed in terms of link flows $\boldsymbol{f}^*$ and of the total non-additive cost $G^{NA}*$ corresponding to deterministic equilibrium:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f}-\boldsymbol{f}^*) + (G^{NA}-G^{NA}*) \geq 0 \quad \forall \boldsymbol{f}=\Delta\boldsymbol{h},\ G^{NA}=(\boldsymbol{g}^{NA})^T\boldsymbol{h} \quad \forall \boldsymbol{h} \in S_h \tag{5.4.5b}$$

An example of Deterministic User Equilibrium assignment for a two-link/path network is shown in Fig. 5.4.3. Note that the deterministic equilibrium flows correspond to the intersection point of the supply and demand curves (in this case, step curves) and they correspond to costs that are equal for the two paths since both are used.

Conditions ensuring the existence and uniqueness of deterministic equilibrium link flows and costs are similar to those described for stochastic equilibrium. In particular, the continuity and monotonicity of the cost functions guarantee respectively the existence and uniqueness of the solution. It should be noted once again that the existence and uniqueness conditions described are only sufficient; i.e. there may exist non-monotone cost functions that give rise to a unique equilibrium vector.

Existence of Deterministic User Equilibrium link flows. The variational inequalities (5.4.3-5) have at least one solution if the cost functions are continuous functions, defined on the non-empty, compact and convex set of the feasible path flows, S_h, or link flows, S_f.

In fact, a general property of variational inequalities is verified, which can be proved through the Brouwer's theorem (see Appendix A).

The considerations regarding the continuity of cost functions made for *SUE* models apply also for DUE models. The existence of equilibrium link flows ensures

the existence of the corresponding link costs, $\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$, and of path costs and flows, $\boldsymbol{g}^*$ and $\boldsymbol{h}^*$, given by the expressions reported in Section 5.2.

Uniqueness of Deterministic User Equilibrium link flows. The variational inequality (5.4.5b), which expresses deterministic equilibrium in terms of link flows, has at most one solution if the link cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are strictly increasing with respect to link flows:

$$[\boldsymbol{c}(\boldsymbol{f}') - \boldsymbol{c}(\boldsymbol{f}'')]^T (\boldsymbol{f}' - \boldsymbol{f}'') > 0 \quad \forall \boldsymbol{f}' \neq \boldsymbol{f}'' \in S_f$$

The same result holds for the variational inequality (5.4.4), which is a special case of (5.4.5b) when non-additive costs are zero.

The proof is by *reductio ad absurdum*. Assume that there exists two different equilibrium link flow vector $\boldsymbol{f}_1^* \neq \boldsymbol{f}_2^* \in S_f$, corresponding to two different feasible path flows vectors, $\boldsymbol{h}_1^* \neq \boldsymbol{h}_2^* \in S_F$, and $G^{NA}{}_1^* = (\boldsymbol{g}^{NA})^T \boldsymbol{h}_1$ and $G^{NA}{}_2^* = (\boldsymbol{g}^{NA})^T \boldsymbol{h}_2^*$ are the relative values of total non-additive cost . Since $\boldsymbol{f}_1^*$ is an equilibrium flow vector, $\boldsymbol{f}_1^*$ and G^{NA^*} must satisfy (5.4.5b) and therefore assuming $\boldsymbol{f} = \boldsymbol{f}_2^* \in S_f$ and $G^{NA} = G^{NA}{}_2^*$ yields:

$$\boldsymbol{c}(\boldsymbol{f}_1^*)^T (\boldsymbol{f}_2^* - \boldsymbol{f}_1^*) + (G^{NA}{}_2^* - G^{NA}{}_1^*) \geq 0$$

Furthermore, $\boldsymbol{f}_2^*$ and $G^{NA}{}_2^*$ also must satisfy (5.4.5b) and therefore assuming $\boldsymbol{f} = \boldsymbol{f}_1^* \in S_f$ and $G^{NA} = G^{NA}{}_1^*$ yields:

$$\boldsymbol{c}(\boldsymbol{f}_2^*)^T (\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) + (G^{NA}{}_1^* - G^{NA}{}_2^*) \geq 0$$

The sum of the two above relationships yields:

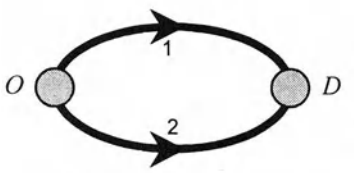
$$\boldsymbol{c}(\boldsymbol{f}_1^*)^T (\boldsymbol{f}_2^* - \boldsymbol{f}_1^*) + \boldsymbol{c}(\boldsymbol{f}_2^*)^T (\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) \geq 0$$

or

$$[\boldsymbol{c}(\boldsymbol{f}_1^*) - \boldsymbol{c}(\boldsymbol{f}_2^*)]^T (\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) \leq 0$$

which contradicts the monotonicity of the cost functions.

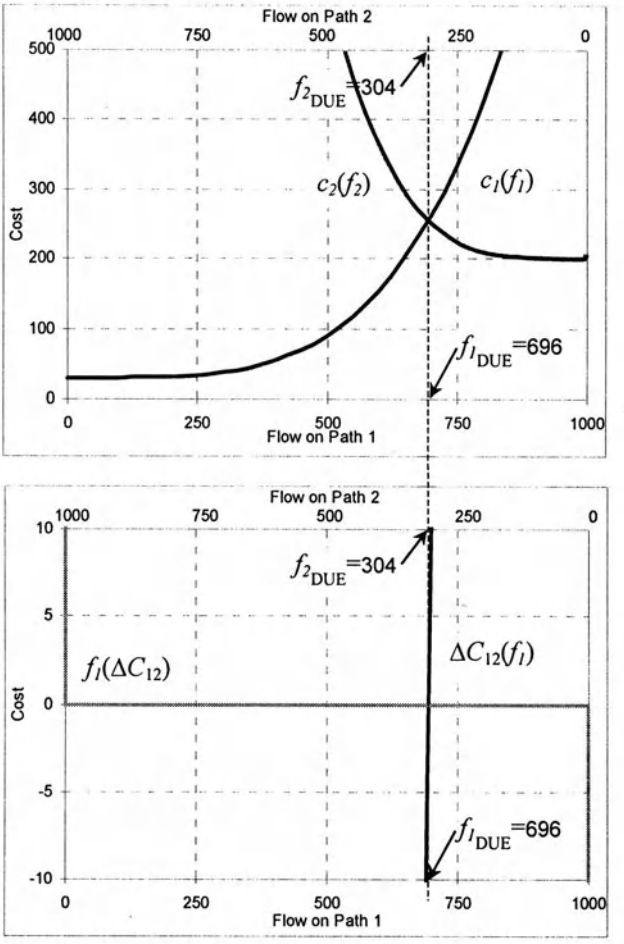
The considerations regarding the monotonicity of the cost functions already made for stochastic equilibrium hold also for the deterministic model. Moreover, the uniqueness of the equilibrium link flows ensures the uniqueness of the corresponding equilibrium link and path costs, $\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$ and $\boldsymbol{g}^* = \boldsymbol{\Delta}^T \boldsymbol{c}^* + \boldsymbol{g}^{NA}$. In general, however, uniqueness of link flows, and therefore of link and path costs, does not ensure the uniqueness of path flows, since there might exist different path flow vectors that induce the same link flows vector, $\boldsymbol{f}^*$, and that correspond to the equilibrium costs, $\boldsymbol{c}^*$ and $\boldsymbol{g}^*$.



$$c_1 = c_{0,1}\left(1 + a_1 \frac{f_1}{Cap_1}\right)^{\gamma} = 30\left(1 + 2\left(\frac{f_1}{500}\right)^4\right);$$

$$c_2 = c_{0,1}\left(1 + a_2 \frac{f_2}{Cap_2}\right)^{\gamma} = 200\left(1 + 2\left(\frac{f_2}{500}\right)^4\right);$$

$$F_1 + F_2 = d_{od} = 1000;\quad F_1 = f_1;\quad F_2 = f_2;$$



Supply equation $\Delta C_{1,2}(f_1) = C_1(f_1) - C_2(f_2=d-f_1)$

Demand equation

$$f_1(\Delta C_{1,2}) = \begin{cases} 0 & \text{if} \quad \Delta C_{1,2} > 0 \\ \in [0, d_{1,2}] & \text{if} \quad \Delta C_{1,2} = 0 \\ d_{1,2} & \text{if} \quad \Delta C_{1,2} < 0 \end{cases}$$

$$f_2 = d_{od} - f_1$$

Fig. 5.4.3 – Example of Deterministic User Equilibrium (DUE)

The non-uniqueness of *DUE* path flows is not particularly relevant from the practical point of view when the main objective of the equilibrium analysisis the simulation of link flows. However, in some applications in which knowledge of path flows is useful or necessary (such as the estimation of the O-D flows using traffic counts described in Chapter 8) this characteristic of deterministic equilibrium assignment may give rise to theoretical and/or algorithmic drawbacks.

Formulation with optimization models. Deterministic equilibrium assignment (with rigid demand) can also be formulated through optimization models, under some assumptions on the cost functions, as now described. These models allow the use of simple and efficient solution algorithms (see Chapter 7). In particular, under the assumptions of separable cost functions and absence of non-additive path costs, the deterministic equilibrium is given by:

$$\boldsymbol{f}^* = \underset{\boldsymbol{f} \in S_f}{argmin} \; \Sigma_l \int_0^{f_l} c_l(y_l)\, dy_l \qquad (5.4.6a)$$

Figure 5.4.4 is a graphic illustration of the model (5.4.6a) and the diagram of the function $z(\boldsymbol{f}) = \Sigma_l \int_0^{f_l} c_l(y_l)\, dy_l$, known as the *integral cost* (its relation to the total cost $\boldsymbol{c}(\boldsymbol{f})^T\boldsymbol{f}$ will be analyzed in Section 5.4.4), for the two-link network introduced in Fig. 5.4.3. Note that the point at which the function $z(\boldsymbol{f})$ has a minimum corresponds to the value of the flows for which the path costs are equal, which are the deterministic equilibrium flows (since both the paths are used).

The formulation (5.4.6a) can be extended to the case of non-separable cost functions as long as they have a symmetric Jacobian (separable functions, with diagonal Jacobian, are clearly a special case):

$$\boldsymbol{f}^* = \underset{\boldsymbol{f} \in S_f}{argmin} \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{y})^T\, d\boldsymbol{y} \qquad (5.4.6b)$$

The assumption of cost functions with a symmetric Jacobian is critical for the formulation of the model (5.4.6b), since generally the line integral value depends on the integration path. However, when the Jacobian $\boldsymbol{Jac}[\boldsymbol{c}(.)]$ of the integrand function $\boldsymbol{c}(.)$ is symmetric, the value of the integral does not depend on the integration path according to Green's theorem (since the set is convex)[15]. In this case the integral depends only on the integration extremes. Thus, since the lower extreme is equal to zero, it depends only on link flow. It is worth pointing out that the Jacobian of non-separable cost functions is rarely symmetric since the way in which the flow on link i affects the cost of link j is generally different from the way in which the flow on link j affects the cost on link i.

The relationship between a solution $\boldsymbol{f}^*$ of the constrained optimization model (5.4.6) and an equilibrium vector can be analyzed by verifying the relationship with

a solution of the variational inequality (5.4.4), as shown below (the demonstrations refer to general features of the variational inequalities; see Appendix A)[16].

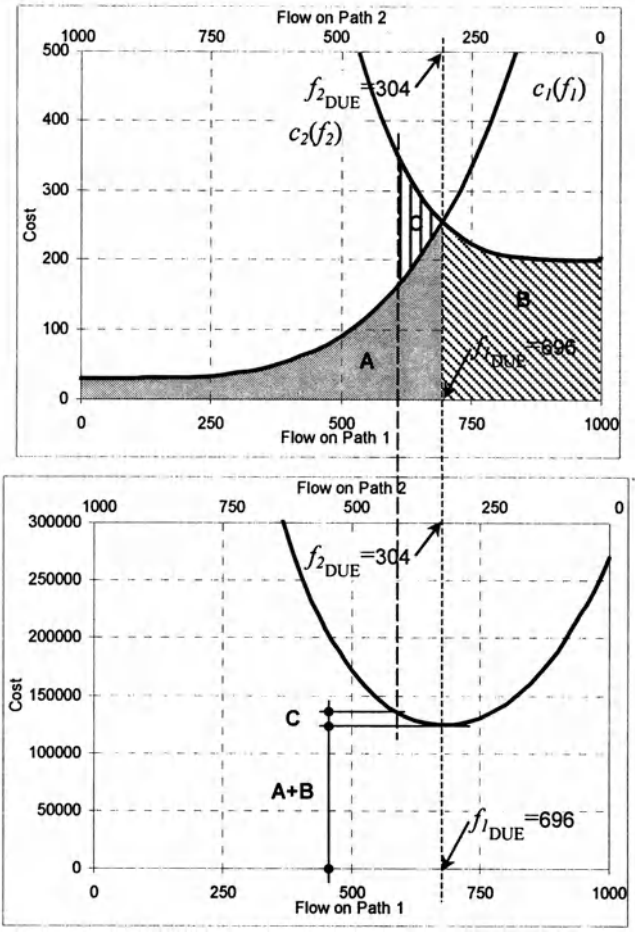


Fig. 5.4.4 – Example of an optimization model for the DUE flows of Fig. 5.4.3.

Equivalence of optimization model for DUE. If the cost functions $\boldsymbol{c}(\boldsymbol{f})$ are continuous with continuous first partial derivative and a symmetric Jacobian, a vector $\boldsymbol{f}^*$ solving the optimization model (5.4.6) is an equilibrium vector (but not necessarily vice versa).

The function $z(\boldsymbol{f}) \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{y})^T \, d\boldsymbol{y}$ for link flow is differentiable with a continuous gradient, since = $\nabla z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})$, and therefore its minimum points satisfy the necessary condition for a minimum (see Appendix A):

$$\nabla z(\boldsymbol{f}^*)^T (\boldsymbol{f} - \boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f$$

Since $\nabla z(\boldsymbol{f}^*) = \boldsymbol{c}(\boldsymbol{f}^*)$, (5.4.4) holds. Furthermore, the function $z(\boldsymbol{f})$ is differentiable, and therefore continuous, on a compact (and convex) set, and therefore has at least one minimum point, consistent with the existence conditions of the solutions of (5.4.4).

If the cost functions $\boldsymbol{c}(\boldsymbol{f})$ are continuous and with continuous first partial derivatives and a symmetric positive semi-definite Jacobian $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$, a vector $\boldsymbol{f}^*$ solving the fixed optimization model (5.4.6) is an equilibrium vector, and vice versa.

Under the above assumptions, $z(\boldsymbol{f})$ is differentiable with a continuous gradient and a continuous positive semi-definite Hessian matrix, since $\nabla z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})$, and $\boldsymbol{Hess}[z(\boldsymbol{f})] = \nabla^2 z(\boldsymbol{f}) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$. Therefore $z(\boldsymbol{f})$ is convex, and its minimum points $\boldsymbol{f}^*$ are defined by the necessary and sufficient condition (see Appendix A):

$$\nabla z(\boldsymbol{f}^*)^T (\boldsymbol{f} - \boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f$$

Since $\nabla z(\boldsymbol{f}^*) = \boldsymbol{c}(\boldsymbol{f}^*)$, (5.4.4) holds. (Furthermore, $z(\boldsymbol{f})$ is convex on a convex set, and therefore has at least one minimum point, consistent with the existence conditions of the solutions of (5.4.4)).

If non-additive path costs differ from zero, the optimization model becomes:

$$\begin{aligned} \boldsymbol{f}^*, G^{NA}* \; = \; & argmin \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{y})^T \, d\boldsymbol{y} + G^{NA} \\ & \boldsymbol{f} = \boldsymbol{\Delta h} \\ & G^{NA} = (\boldsymbol{g}^{NA})^T \boldsymbol{h} \\ & \boldsymbol{h} \in S_h \end{aligned} \tag{5.4.7}$$

The model (5.4.7) has properties analogous to those shown above for model (5.4.6b).

When the cost function $\boldsymbol{c}(\boldsymbol{f})$ has a symmetric positive definite Jacobian, the objective functions of models (5.4.6) and (5.4.7) respectively have a single minimum point (unimodal functions). In particular, the objective function of model (5.4.6) is strictly convex and therefore has a single minimum point, consistent with the uniqueness conditions presented with the variational inequality models, since under this assumption the cost functions are strictly increasing. However, the objective function of model (5.4.7) is convex with a single minimum point, since it is the summation of a function which is strictly convex with respect to the variables $\boldsymbol{f}$ and a linear function with respect to the variable G^{NA}.

Deterministic user equilibrium link flows can be calculated with various algorithms solving the variational inequality or optimization models in the case of cost functions with symmetric Jacobian[17]. Some simple algorithms which use deterministic network loading assignment are described in Chapter 7.

5.4.3. Relationship between stochastic and deterministic equilibrium flows

The deterministic path choice model underlying deterministic equilibrium models can be considered a special case of a random utility model in which the variance of the random residuals is null. For this reason, stochastic equilibrium flows approximate deterministic equilibrium flows as the random residual variance goes to zero. Figure 5.4.5 shows the curves expressing the demand model for the example used in Figures 5.4.2 -.3, for various values of the parameter θ proportional to the standard deviation of random residuals of the path choice model. The figure clearly shows that the probabilistic demand curve progressively approaches the curve corresponding to the deterministic model and SUE flows approach DUE flows.

Deterministic and stochastic models give similar results in the case of very congested networks. If link flows are close to capacity, the derivatives of the cost functions, representing the cost variations introduced by marginal user, are most likely larger than the random residuals. In other words, a flow distribution very different from deterministic equilibrium would induce large cost differences between the different paths that are likely to be correctly perceived by almost all the users.

This effect is shown in Fig. 5.4.6, where the link cost functions vary in such a way that their derivatives increase but their traversing point remains fixed. In other words, DUE flows remain unchanged while the system is more congested, and thus more sensitive to small flow variations. As the figure shows, as the cost curves vary, SUE flows change and approach DUE flows.

The closeness of deterministic and stochastic equilibrium flows implies that on very congested networks it is possible to use DUE assignment as an approximation of SUE assignment. This is good for practical problems, since DUE flows are easier to compute, as will be shown in Chapter 7. However, it should be noted that for different applications (assignment to lightly congested or not uniformly congested networks, estimation of the O-D matrix by traffic counts, etc.) the deterministic model is not a good substitute for the stochastic one. Furthermore, as pointed out in the preceding section, it is not possible to guarantee the uniqueness of deterministic equilibrium path flows, nor (as will be seen in Section 5.7) of flows per user class in the case of multi-class assignment.

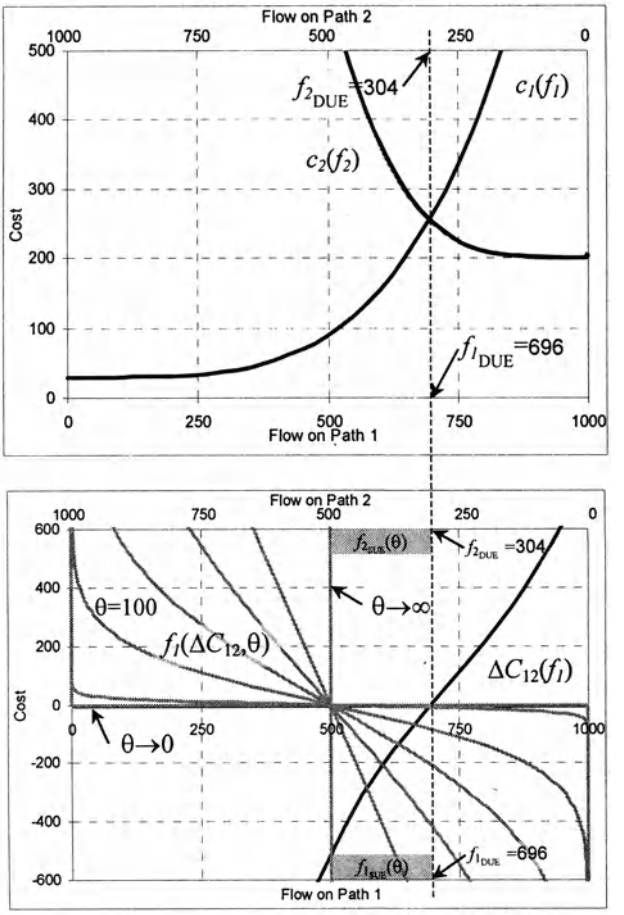
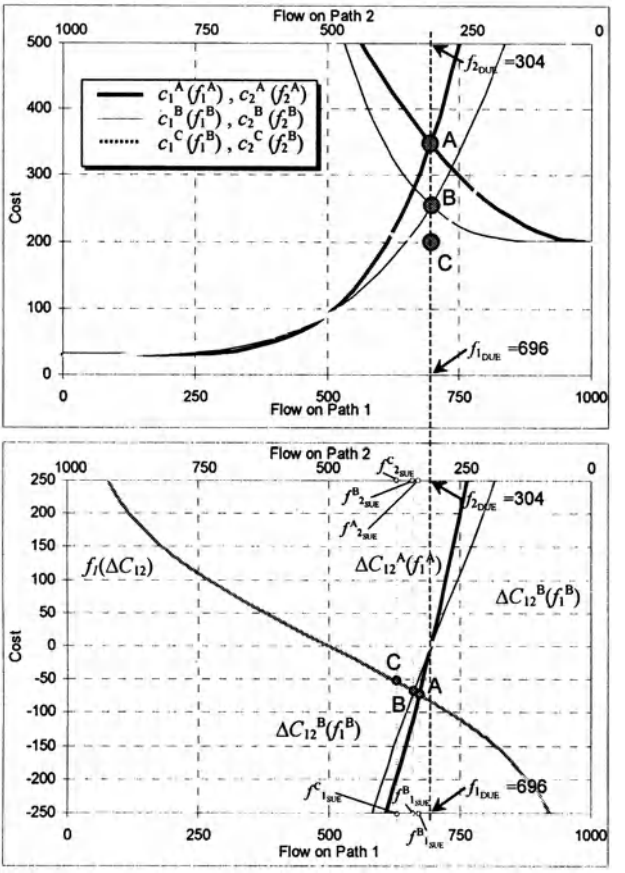


Fig. 5.4.5 Relationship between SUE and DUE flows for different values of random residuals standard deviation (see Fig. 5.4.2 and Fig. 5.4.3).



Cost functions

$$c_1 = c_{0,1}\left(1 + a_1 \frac{f_1}{Cap_1}\right)^{\gamma_1};$$

$$c_2 = c_{0,2}\left(1 + a_2 \frac{f_2}{Cap_2}\right)^{\gamma_2};$$

Parameters of the cost functions

Case	$c_{0,1}$	$c_{0,2}$	a_1	a_2	Cap_1	Cap_2	γ_1	γ_2
A	30	200	2	2	500	500	5.06	2
B	30	200	2	2	500	500	4	4
C	30	200	2	2	500	500	3.17	12

θ=100

Fig. 5.4.6 – Relationship between SUE and DUE flows for varying cost functions.

5.4.4. System optimal assignment models

System optimal assignment models derive from assumptions which are significantly different from those underlying the user equilibrium models. In fact, it is assumed that users "cooperate" to minimize total system cost, rather than minimizing individual costs as in user equilibrium models. The resulting assignment problem is generally different, at least for congested networks; it is equal to deterministic assignment for uncongested networks. Note that under the assumptions of system optimal assignment, some users may follow a non-minimum (perceived or systematic) cost path. This condition, under which "the total cost on the network is minimum" as expressed by the *Second principle of Wardrop*, is known as System Optimum (SO).

Knowledge of system optimum flows can be useful as a reference element in the analysis of congested networks. In fact, although the behavioral assumptions underlying SO are not realistic for the simulation of individual decision-maker behavior, the system optimum flows and costs correspond to (one of) the system "management" objectives to be achieved through available control instruments (prices, traffic-light regulation, services frequency, etc.)[18]. Furthermore, SO assignment can be applied for the assignment of demand units without autonomous decision capability, such as freight vehicles.

SO assignment is defined by an optimization model in terms of link flows with an objective function consisting of the total cost presented in Section 5.2 (ignoring non-additive path costs for the sake of simplicity):

$$\begin{aligned} \boldsymbol{f}_{SO} &= \operatorname{argmin}\ \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f} \\ \boldsymbol{f} &\in S_f \end{aligned} \tag{5.4.8}$$

Note that it is unnecessary to introduce assumptions on the symmetry of the cost function Jacobian to formulate system optimum assignment through an optimization model (which in this case is the direct formulation, rather than an equivalent indirect one as with DUE). The existence and uniqueness of optimum system flows and costs are discussed below.

Existence. The optimization model (5.4.8) has at least one solution if the cost functions, $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$, are continuous.

> Under these assumptions, the objective function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$, is continuous on the non-empty (under the assumption of connected network), and compact (as well as convex) set S_f, and therefore has at least one minimum point (see Appendix A).

Existence and uniqueness. The optimization model (5.4.8) has one and only one solution if the cost functions, $\boldsymbol{c}(\boldsymbol{f})$, have continuous first and second partial

derivatives, their Jacobian, $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$, is continuous and positive definite (cost functions are strictly increasing), and the Hessian matrix, $\boldsymbol{Hess}[c_l(\boldsymbol{f})]$, of each cost function, $c_l = c_l(\boldsymbol{f})$, positive semi-definite (each cost function is convex).

Under these assumptions, the cost functions have continuous first derivatives and are therefore differentiable and continuous, a condition guaranteeing the existence of the solution. Furthermore, the gradient, $\nabla z(\boldsymbol{f})$, of the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$ is given by:

$$\nabla z(\boldsymbol{f}) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]\boldsymbol{f} + \boldsymbol{c}(\boldsymbol{f})$$

and the Hessian matrix, $\boldsymbol{Hess}[z(\boldsymbol{f})]$, of the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$ is given by:

$$\boldsymbol{Hess}[z(\boldsymbol{f})] = \boldsymbol{Jac}[\nabla z(\boldsymbol{f})] = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]^T + \Sigma_l f_l \boldsymbol{Hess}[c_l(\boldsymbol{f})] + \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$$

Both $z(\boldsymbol{f})$ and $\boldsymbol{Hess}\,[z(\boldsymbol{f})]$ are continuous, so that the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$ is twice differentiable. Finally, the Hessian matrix, $\boldsymbol{Hess}[z(\boldsymbol{f})]$, is symmetric positive definite since it is the sum of symmetric positive semi-definite matrices and of symmetric positive definite matrices. Therefore, the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$, defined over the convex S_f, is strictly convex and has one and only one minimum point.

System optimum flows do not generally coincide with DUE flows, as is shown by the example reported in Fig. 5.4.7. The figure shows that, with respect to the DUE flows, the shift of some users on a path slightly more expensive but less congested reduces significantly the total cost borne by all users.

However, if link costs are independent of flows, that is $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})] = 0$, the solutions to the two problems coincide.

If $\boldsymbol{f}^*$ is a minimum point of the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$ it follows that:

$$\nabla z(\boldsymbol{f}^*)^T (\boldsymbol{f} - \boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f$$

and since $\nabla z(\boldsymbol{f}^*) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]\boldsymbol{f} + \boldsymbol{c}(\boldsymbol{f})$:

$$(\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]\boldsymbol{f} + \boldsymbol{c}(\boldsymbol{f}))^T (\boldsymbol{f} - \boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f$$

a condition which in general is different from the variational inequality (5.4.6) which expresses the deterministic equilibrium. However if link costs are independent of the flows ($\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})] = 0$) the above inequality coincides with the variational inequality, the deterministic user equilibrium is reduced to the deterministic uncongested network assignment and can be expressed by the model (5.3.7) equivalent in this case to the model (5.4.8) expressing the system optimum.

Of particular interest is the example in Fig. 5.4.8, known in the literature as Braess paradox. The paradox refers to a network where the addition of a new link causes an increase of the total cost under the deterministic equilibrium assignment, while leaving the system optimum total cost unchanged. In the first case, the SO link flows minimize the total cost and for this reason the addition of a link cannot increase the overall cost of the system because the SO link flow pattern corresponding to the null flow on the new link is a feasible solution of the new SO problem. Vice versa, in the case of user equilibrium, the objective of each individual is to minimize his or her own transport cost and the equilibrium link flow pattern corresponding to the introduction of a new link may cause an increase of total cost. It should be pointed out, however, that conditions analogous to the Braess paradox are not often found in real systems[19].

The system optimum model (5.4.8) can be reformulated to be formally analogous to the DUE optimization model (5.4.6). To do so, consider the vector function $\boldsymbol{b}(\boldsymbol{f})$ of the link flow vector defined by the gradient $\nabla z(\boldsymbol{f})$ of the function $z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$ and called the marginal cost function:

$$\boldsymbol{b}(\boldsymbol{f}) = \nabla z(\boldsymbol{f}) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]^{\mathrm{T}} \boldsymbol{f} + \boldsymbol{c}(\boldsymbol{f}) \tag{5.4.9}$$

The interpretation of the function (5.4.9) is more apparent in the case of separable cost functions $\boldsymbol{c}(\boldsymbol{f})$ where the functions $\boldsymbol{b}(\boldsymbol{f})$ are also separable. Under this assumption, if the first derivative of the link l cost function $c_l(f_l)$ is denoted by $c'_l(f_l)$, it follows that:

$$b_l(f_l) = c'_l(f_l) f_l + c_l(f_l)$$

In the general case, if the cost functions have continuous first and second derivatives, the Jacobian $\boldsymbol{Jac}[\boldsymbol{b}(\boldsymbol{f})]$ of the function $\boldsymbol{b}(\boldsymbol{f})$ is symmetric and therefore the line integral of $\boldsymbol{b}(\boldsymbol{f})$ between the values $\boldsymbol{0}$ and $\boldsymbol{f}$ does not depend on the integration path and its value coincides with the total cost:

$$\int_{\boldsymbol{0}}^{\boldsymbol{f}} \boldsymbol{b}(\boldsymbol{y})^{\mathrm{T}} \, d\boldsymbol{y} = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$$

If the cost functions, $c(G+I)$, have continuous first and second derivatives, the marginal costs, $\boldsymbol{b}(\boldsymbol{f})$, have continuous cost derivatives, thus they are differentable and continuous. Furthermore, the Jacobian, $\boldsymbol{Jac}[\boldsymbol{b}(\boldsymbol{f})]$, of the gradient function, $\boldsymbol{b}(\boldsymbol{f}) = \nabla z(\boldsymbol{f})$, coinciding with the Hessian matrix of the function, $z(\boldsymbol{f})$, is symmetric.

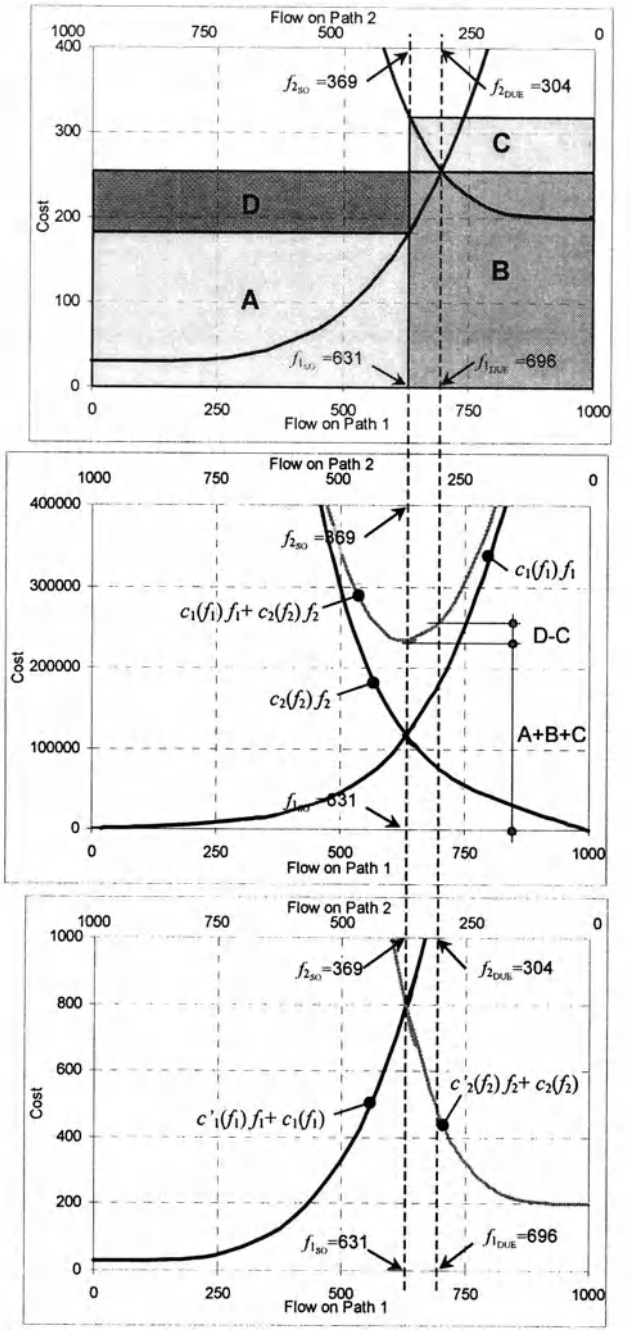
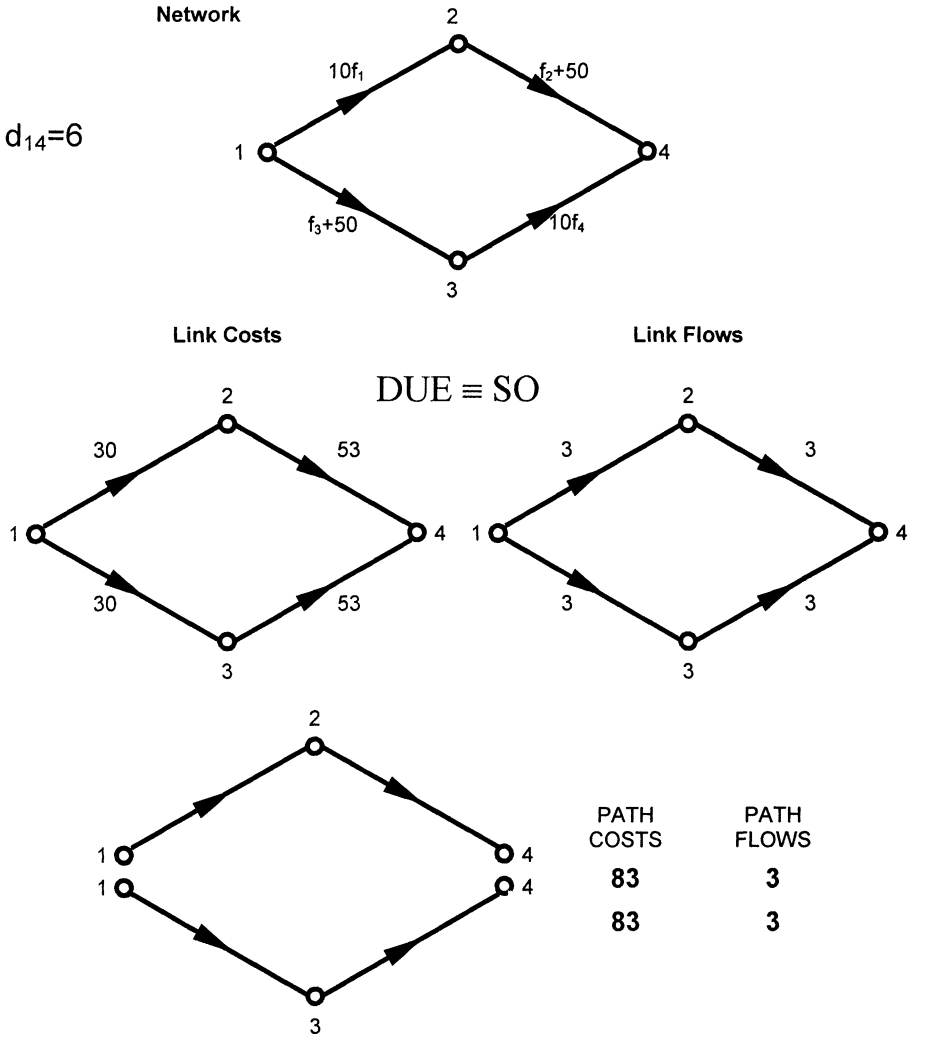


Fig. 5.4.7 System optimum (SO) flows on the test network of Fig. 5.4.3.



g_{min} = 83

TC = 83·3+83·3 = 498

Fig. 5.4.8 (a) – Example of the Braess' paradox

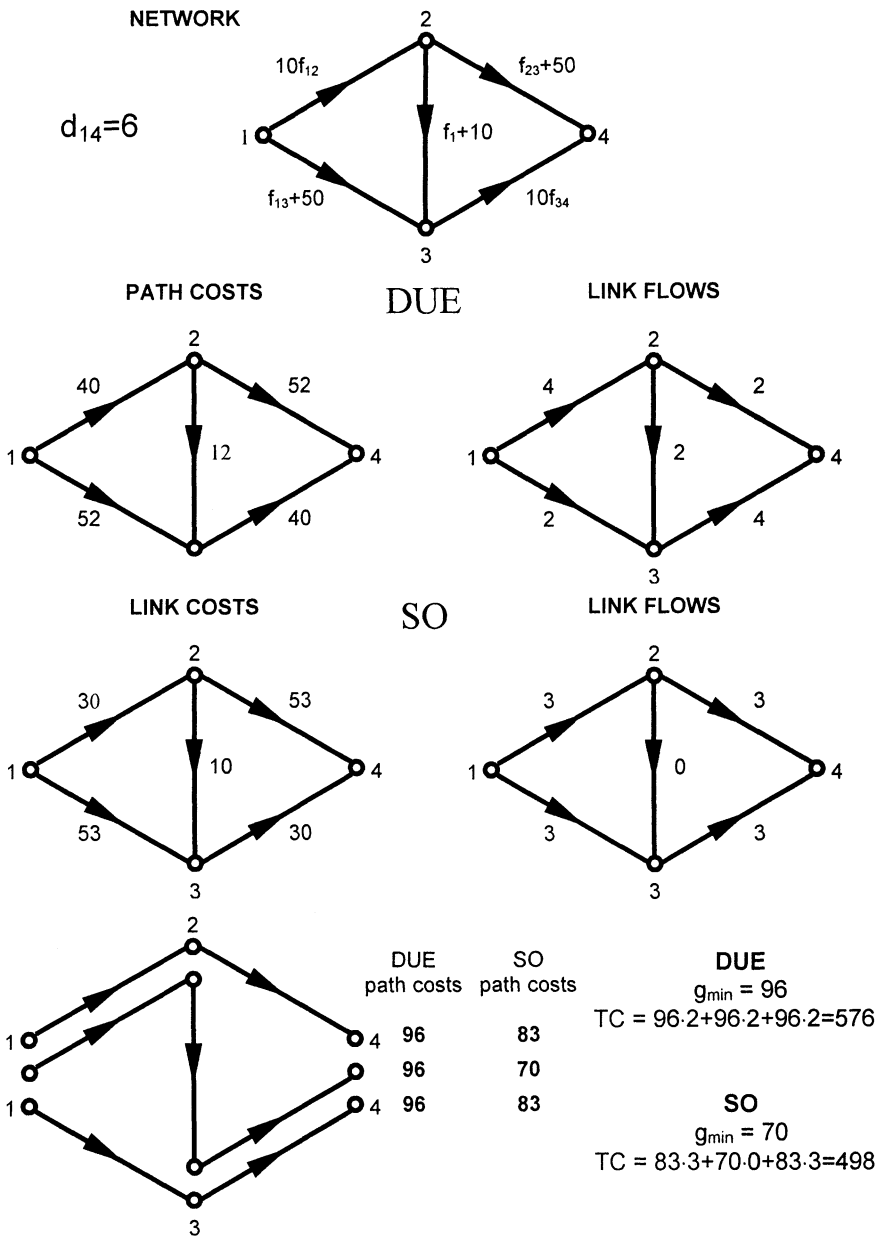


Fig. 5.4.8 (b) – Example of the Braess' paradox

System optimum assignment can therefore be formulated with an optimization model by using the marginal cost function $\boldsymbol{b}(\boldsymbol{y})$ defined by (5.4.9):

$$\boldsymbol{f}_{SO} = \underset{\boldsymbol{f} \in S_f}{argmin}\, z(\boldsymbol{f}) = \int_0^{\boldsymbol{f}} \boldsymbol{b}(\boldsymbol{y})\, d\boldsymbol{y} \qquad (5.4.10)$$

The optimization model (5.4.10) is formally analogous to the optimization model (5.4.6) for the (symmetric) DUE and can be solved with the same algorithms described in Chapter 7 (see Fig. 5.4.6).

An operational interpretation of the model (5.4.10) is that if link costs are modified in such a way that the link costs perceived by the users' coincide with the marginal ones, $\boldsymbol{b}(\boldsymbol{f})$, individual deterministic path choice with respect to such costs would lead to a configuration for the entire system which minimized the total cost $\boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$. A way (but not the only one) of introducing this variation of costs is the application of efficiency tolls as a function of the link flows given by: $\boldsymbol{b}(\boldsymbol{f}) - \boldsymbol{c}(\boldsymbol{f}) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]\boldsymbol{f}$. In the case of separable costs, this expression is reduced to: $c'_l(f_l)\, f_l$. This point will be returned to in Chapter 9 in the discussion of supply design models.

Finally, its is possible to deduce that system optimum assignment does not generally coincide with stochastic equilibrium through similar arguments. In this case, it is also possible to derive equivalence conditions which lead to rather unrealistic cost functions.

5.5. Assignment models with pre-trip/en-route path choice

The previous sections deal with the case in which users, before starting the trip, choose between single paths representing routes provided by the transport system. The analysis can, however, quite easily be extended to include pre-trip/en-route path choice behavior, relevant in public transport systems with high frequency and/or low reliability. In this case (as was seen in Section 4.3.4.2) the relevant pre-trip choice alternatives are en-route strategies, whose topology is represented by network hyperpaths, while en-route choices are made during the trip itself at each diversion (waiting) node where different lines are available (Fig. 5.5.1). In any case, the approach described in this section can be applied to other transportation systems once en-route diversion nodes and the related choice behavior has been specified.

The main modifications surround the demand model defined in Section 5.2 by equation (5.2.7). In particular, the difference between path and hyperpath costs and flows is defined by the path choice probabilities within the hyperpaths. Referring to notation introduced in Section 4.2.5.2, (see Fig. 5.5.1). Let:

$\omega_{od,kj}$ be the conditional probability of choosing path k within the hyperpath j for a user of the od pair;

Ω_{od} be the matrix of conditional path choice probabilities $\omega_{od,kj}$ within the hyperpaths for the od pair;

Ω be the overall matrix of conditional path choice probabilities for all paths, all hyperpaths and all od pairs, obtained by placing the blocks Ω_{od} corresponding to each od pair side by side.

By analogy with the path definitions in Section 5.2, additive and non-additive costs are taken into consideration for each hyperpath. Let:

$\boldsymbol{x}_{od}^{ADD}$ be the hyperpath additive costs vector for the users of the od pair;

$\boldsymbol{x}^{ADD}$ be the overall vector of the hyperpath additive costs, consisting of the hyperpath additive cost vectors $\boldsymbol{x}_{od}^{ADD}$ corresponding to each od pair;

$\boldsymbol{x}_{od}^{NA}$ be the hyperpath non-additive costs vector for the users of the od pair;

$\boldsymbol{x}^{NA}$ be the overall vector of the hyperpath non-additive costs, consisting of the hyperpath non-additive cost vectors $\boldsymbol{x}_{od}^{NA}$ corresponding to each od pair;

$\boldsymbol{x}_{od}$ be the vector of the total hyperpath costs for the users of the od pair;

$\boldsymbol{x}$ be the overall vector of the total hyperpath costs, consisting of the vectors of the total hyperpath costs vectors $\boldsymbol{x}_{od}$ corresponding to each od pair.

As was seen in Section 4.2.5.2, the hyperpath additive costs $\boldsymbol{x}_{od}^{ADD}$ are usually defined by a linear combination of on-board time $\boldsymbol{Tb}$, as well as by the access/egress times $\boldsymbol{Ta}$ and by the boarding and alighting times $\boldsymbol{Tbr}$ and $\boldsymbol{Tal}$ homogenized with suitable coefficients:

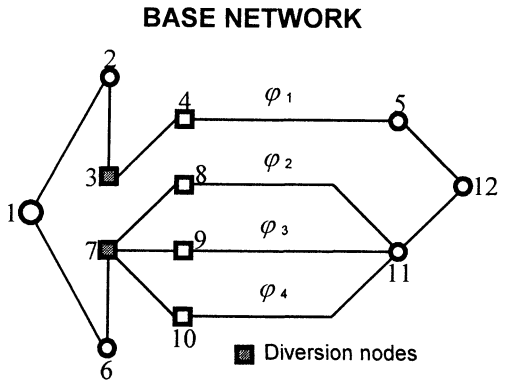
$$\boldsymbol{x}_{od}^{ADD} = \beta_b \boldsymbol{Tb} + \beta_{br} \boldsymbol{Tbr} + \beta_{al} \boldsymbol{Tal} + \beta_d \boldsymbol{Td} + \beta_a \boldsymbol{Ta} \qquad \forall od$$

Furthernore, non-additive hyperpath costs, $\boldsymbol{x}_{od}^{NA}$, usually include performance attributes which cannot be computed from generic link costs such as the waiting time $\boldsymbol{Tw}_{od}$ and the number of transfers, $\boldsymbol{N}_{od}$, homogenized with suitable coefficients:

$$\boldsymbol{x}_{od}^{NA} = \beta_w \boldsymbol{Tw}_{od} + \beta_N \boldsymbol{N}_{od} \quad \forall od$$

The relationship between the hyperpath costs and the additive path costs is expressed by the following:

$$\boldsymbol{x}_{od}^{ADD} = \boldsymbol{\Omega}_{od}^{T} \boldsymbol{g}_{od}^{ADD} \qquad \forall od$$



Ω Matrix
(nr. of path × nr. of hyperpath

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.4 & 0.44 & 0 & 0.27 \\ 0 & 0 & 1 & 0 & 0.6 & 0 & 0.55 & 0.40 \\ 0 & 0 & 0 & 1 & 0 & 0.56 & 0.45 & 0.33 \end{bmatrix}$$

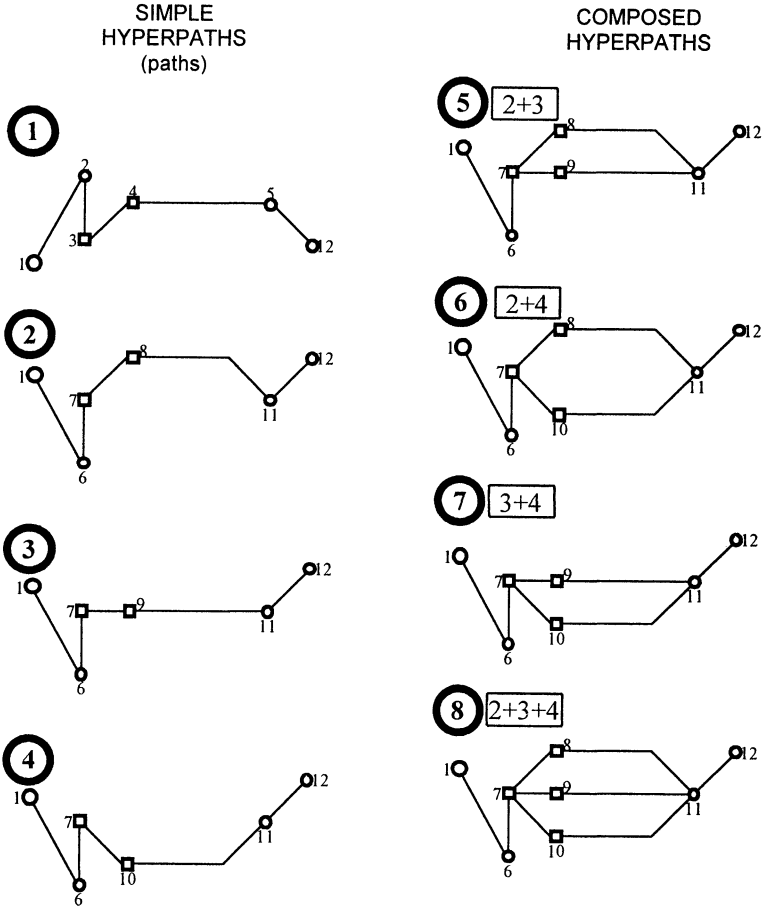


Fig. 5.5.1 Example of conditional path choice matrix.

In the following, for simplicity of notation, it is assumed that any non-additive path costs $\boldsymbol{g}^{NA}$ have been included in the non-additive hyperpath costs $\boldsymbol{x}^{NA}$, and therefore the path costs $\boldsymbol{g}$ coincide with the additive costs $\boldsymbol{g}^{ADD}$ (Fig. 5.5.2)

$$\boldsymbol{x}_{od} = \boldsymbol{\Omega}_{od}^{T}\boldsymbol{g}_{od} + \boldsymbol{x}_{od}^{NA} \qquad \forall od \qquad (5.5.1a)$$

$$\boldsymbol{x} = \boldsymbol{\Omega}^{T}\boldsymbol{g} + \boldsymbol{x}^{NA} \qquad (5.5.1b)$$

$$\boldsymbol{x} = \boldsymbol{\Omega}^{T} \bullet \boldsymbol{g}^{ADD} + \boldsymbol{x}^{NA}$$

$$\begin{bmatrix} 471 \\ 461 \\ 481 \\ 816 \\ 474 \\ 667 \\ 630 \\ 594 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.40 & 0.60 & 0 \\ 0 & 0.44 & 0 & 0.56 \\ 0 & 0 & 0.55 & 0.45 \\ 0 & 0.27 & 0.40 & 0.33 \end{bmatrix} \bullet \begin{bmatrix} 421 \\ 421 \\ 451 \\ 771 \end{bmatrix} + \begin{bmatrix} 50 \\ 40 \\ 30 \\ 45 \\ 35 \\ 50 \\ 35 \\ 45 \end{bmatrix}$$

Fig. 5.5.2 Relationship between path and hyperpath costs for the network of Fig. 5.5.1

The choice of strategy, i.e. of the hyperpath representing its topology, is simulated by a random utility model where the systematic utility of a hyperpath is equal to the opposite of the hyperpath systematic cost, analogously to (5.2.5) (Section 4.2.5.2):

$$\boldsymbol{V}_{od} = -\boldsymbol{x}_{od} + \boldsymbol{V}^{\circ}_{od} = -\boldsymbol{\Omega}_{od}^{T}\boldsymbol{g}_{od}^{ADD} - \boldsymbol{x}_{od}^{NA} + \boldsymbol{V}^{\circ}_{od} \qquad \forall od \qquad (5.5.2)$$

where:

$\boldsymbol{V}_{od}$ is a vector with an element for each hyperpath j, given by the systematic utility V_j of the hyperpath j, for the users of the od pair;

$\boldsymbol{V}^{\circ}_{od}$ is a vector with elements consisting of the sum of any other attributes which cannot be assigned to hyperpath costs (such as socio-economic attributes of the users), from now on ignored for simplicity of notation.

The hyperpath choice probabilities depend on the systematic utilities of the hyperpaths, and therefore on the systematic costs. Let[20]:

$q[j/od]$ be the probability that a user, during a trip from the origin o to the destination d (with purpose, time band and mode not explicitly indicated), uses the hyperpath j;

$\boldsymbol{q}_{od}$ be the vector of the hyperpath choice probabilities for the users of the od pair, whose elements are the probabilities $q[j/od]$ with hyperpath index j varying within the set of all hyperpaths; this set is assumed non empty (each od pair is connected by at least one hyperpath) and finite (only elementary hyperpaths are considered).

As shown in Section 4.2.5.2, hyperpath choice probabilities can be expressed through random utility models as:

$$q[j/od] = Prob[\ V_j - V_{mj} \geq \varepsilon_m - \varepsilon_j \quad \forall\ m] \quad \forall od,j$$
$$\boldsymbol{q}_{od} = \boldsymbol{q}_{od}(\boldsymbol{V}_{od}) \quad \forall od$$

where ε_j is the random residual corresponding to the perceived utility of hyperpath j.

Combining the hyperpath choice model with the systematic utility specification, the relation between hyperpath choice probabilities and costs for the od pair, known as the *hyperpath choice map*, is obtained:

$$\boldsymbol{q}_{od} = \boldsymbol{q}_{od}\,(\boldsymbol{V}_{od} = -\boldsymbol{x}_{od}) \qquad \forall od$$

The above relationship can be expressed in matrix terms. Let:

$\boldsymbol{Q}$ be the hyperpath choice probabilities matrix with a column for each od pair and a row for each hyperpath j, with entries given by $q[j/od]$ if the hyperpath j connects the od pair, zero otherwise (the matrix $\boldsymbol{Q}$ is block diagonal with blocks given by the vectors $\boldsymbol{q}_{od}$).

Therefore

$$\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{V} = -\boldsymbol{x})$$

The (average) flow y_j on the hyperpath j connecting the od pair is given by the product of the corresponding demand flow d_{od} and the corresponding hyperpath choice probability:

$$y_j = d_{od}\, q[j/od]$$

and is measured in the demand units. Let

$\boldsymbol{y}_{od}$ be the hyperpath flow vector for the *od* pair whose elements are the flows y_j, with hyperpath index j varying within the set of hyperpaths.

$\boldsymbol{y}$ be the overall vector of hyperpath flows, consisting of the vectors of the hyperpath flows $\boldsymbol{y}_{od}$ corresponding to each *od* pair.

For each *od* pair, the relation between hyperpath choice probabilities and flows and demand flows, parallel to (5.2.6), is expressed by:

$$\boldsymbol{y}_{od} = d_{od}\, \boldsymbol{q}_{od}(\boldsymbol{V}_{od}) \quad \forall od \tag{5.5.3a}$$

$$\boldsymbol{y} = \boldsymbol{Q}(\boldsymbol{V})\boldsymbol{d} \tag{5.5.3b}$$

Each path k which connects the *od* pair may belong to several hyperpaths, so that the flow h_k is given by the sum of the hyperpath flows y_j for the probability $\omega_{od,kj}$ that the path k is used within the hyperpath j (an example is reported in Fig. 5.5.3):

$$h_k = \Sigma_h\, \omega_{od,kj}\, y_j \quad \forall k$$

$$\boldsymbol{h}_{od} = \boldsymbol{\Omega}_{od}\, \boldsymbol{y}_{od} = d_{od}\, \boldsymbol{\Omega}_{od}\, \boldsymbol{q}_{,od} \quad \forall od \tag{5.5.4a}$$

$$\boldsymbol{h} = \boldsymbol{\Omega}\boldsymbol{y} = \boldsymbol{\Omega}\boldsymbol{Q}\boldsymbol{d} \tag{5.5.4b}$$

$$\boldsymbol{h} = \boldsymbol{\Omega} \bullet \boldsymbol{y}$$

$$\begin{bmatrix} 138 \\ 244 \\ 283 \\ 214 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.4 & 0.44 & 0 & 0.27 \\ 0 & 0 & 1 & 0 & 0.6 & 0 & 0.55 & 0.4 \\ 0 & 0 & 0 & 1 & 0 & 0.56 & 0.45 & 0.33 \end{bmatrix} \quad \begin{bmatrix} 138 \\ 139 \\ 136 \\ 97 \\ 137 \\ 113 \\ 117 \\ 0 \end{bmatrix}$$

Fig. 5.5.3 Relationship between path and hyperpath flows for the network of Fig. 5.5.1.

The complete demand model in the case of pre-trip/en-route path choice behavior is defined by relations (5.5.1-2) specifying the systematic utility and by the relations (5.5.3-4) defining the path flows. When combined, these relations lead to a relation between path flows and costs which generalizes expression (5.2.7):

$$\boldsymbol{h}_{od} = d_{od}\, \boldsymbol{\Omega}_{od}\, \boldsymbol{q}_{,od}(-\boldsymbol{\Omega}_{od}^{T}\boldsymbol{g}_{od} - \boldsymbol{x}_{od}^{NA}) \;\forall od \tag{5.5.5a}$$

$$\boldsymbol{h} = \boldsymbol{\Omega}\boldsymbol{Q}(-\boldsymbol{\Omega}^{T}\boldsymbol{g} - \boldsymbol{x}^{NA})\,\boldsymbol{d} \tag{5.5.5b}$$

By combining the demand model (5.5.5) with the supply model (5.2.4), the assignment models described in the previous sections can be extended to handle pre-trip/en-route path choice behavior . In this case, it is useful to express the relation between link and hyperpath flows and costs. Let

$\lambda_{od,lj}$ be the the probability of traversing link l within the hyperpath j for the users of the od pair;

Λ_{od} be the matrix of the traversing probabilities $\lambda_{od,lj}$ of each link l within each hyperpath j for the users of the od pair;

Λ be the overall matrix of the probabilities of traversing the links within each hyperpath, consisting of the blocks Λ_{od} corresponding to each od pair.

The relationship between the link traversing probabilities and the path choice probabilities within a hyperpath (analogous to the link-path incidence described in Chapter 2 and repeated in Section 5.2), is expressed by the following relations (fig. 5.5.4):

$$\lambda_{od,lj} = \Sigma_k \ \delta_{od,lk} \ \omega_{od,kj} \quad \forall l \ \forall od$$

$$\Lambda_{od} = \Delta_{od} \Omega_{od} \quad \forall od \qquad (5.5.6a)$$

$$\Lambda = \Delta \ \Omega \qquad (5.5.6b)$$

Λ (links × hyperpaths) =

	1	2	3	4	5	6	7	8
1-2	1	0	0	0	0	0	0	0
1-6	0	1	1	1	1	1	1	1
2-3	1	0	0	0	0	0	0	0
3-4	1	0	0	0	0	0	0	0
4-5	1	0	0	0	0	0	0	0
5-12	0	1	1	1	1	1	1	1
6-7	0	1	1	1	1	1	1	1
7-8	0	1	0	0	0.4	0.44	0	0.27
7-9	0	0	1	0	0.6	0	0.55	0.40
7-10	0	0	1	0	0.6	0	0.55	0.40
8-11	0	0	1	0	0.6	0	0.55	0.40
9-11	0	0	0	1	0	0.56	0.45	0.33
10-11	0	0	0	1	0	0.56	0.45	0.33
11-12	0	1	1	1	1	1	1	1

= Δ (links × paths) • Ω (paths × hyperpaths)

	1	2	3	4
1-2	1	0	0	0
1-6	0	1	1	1
2-3	1	0	0	0
3-4	1	0	0	0
4-5	1	0	0	0
5-12	1	0	0	0
6-7	0	1	1	1
7-8	0	1	0	0
7-9	0	0	1	0
7-10	0	0	0	1
8-11	0	1	0	0
9-11	0	0	1	0
10-11	0	0	0	1
11-12	0	1	1	1

	1	2	3	4	5	6	7	8
1	1	0	0	0	0	0	0	0
2	0	1	0	0	0.4	0.44	0	0.27
3	0	0	1	0	0.6	0	0.55	0.40
4	0	0	0	1	0	0.56	0.45	0.33

Fig. 5.5.4 Incidence and traversing probability matrices for the network of Fig. 5.5.1

The relationship between hyperpath costs, link costs and additive path costs is expressed by the following equation, obtained by combining expressions (5.5.1), (5.5.6) and (5.2.1) (non-additive path costs $\boldsymbol{g}^{NA}$ have been included in the non-additive hyperpath costs $\boldsymbol{x}_{od}^{NA}$, thus path costs coincide with additive costs: $\boldsymbol{g} = \boldsymbol{g}^{ADD}$):

$$\boldsymbol{x}_{od} = \boldsymbol{x}_{od}^{ADD} + \boldsymbol{x}_{od}^{NA} = \boldsymbol{\Omega}_{od}^{T} \boldsymbol{\Delta}_{od}^{T} \boldsymbol{c} + \boldsymbol{x}_{od}^{NA} = \boldsymbol{\Lambda}_{od}^{T} \boldsymbol{c} + \boldsymbol{x}_{od}^{NA} \qquad \forall od \qquad (5.5.7a)$$

$$\boldsymbol{x} = \boldsymbol{x}^{ADD} + \boldsymbol{x}^{NA} = \boldsymbol{\Omega}^{T} \boldsymbol{\Delta}^{T} \boldsymbol{c} + \boldsymbol{G}^{NA} = \boldsymbol{\Lambda}^{T} \boldsymbol{c} + \boldsymbol{x}^{NA} \qquad (5.5.7b)$$

Similarly, the relationship between link and hyperpath flows is expressed by the following equation by combining expressions (5.5.4), (5.5.6) and (5.2.3) (Fig. 5.5.5):

$$\boldsymbol{f} = \Sigma_{od} \boldsymbol{\Delta}_{od} \boldsymbol{h}_{od} = \Sigma_{od} \boldsymbol{\Delta}_{od} \boldsymbol{\Omega}_{od} \boldsymbol{y}_{od} = \Sigma_{od} \boldsymbol{\Lambda}_{od} \boldsymbol{y}_{od} \qquad (5.5.8a)$$

$$\boldsymbol{f} = \boldsymbol{\Delta} \boldsymbol{h} = \boldsymbol{\Delta} \boldsymbol{\Omega} \boldsymbol{y} = \boldsymbol{\Lambda} \boldsymbol{y} \qquad (5.5.8b)$$

$\boldsymbol{f}$ = $\boldsymbol{\Delta}$ • $\boldsymbol{h}$

$$\begin{bmatrix} 126 \\ 874 \\ 126 \\ 126 \\ 126 \\ 126 \\ 874 \\ 254 \\ 356 \\ 264 \\ 254 \\ 356 \\ 264 \\ 874 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \bullet \begin{bmatrix} 126 \\ 254 \\ 356 \\ 264 \end{bmatrix}$$

$\boldsymbol{f}$ = $\boldsymbol{\Lambda}$ • $\boldsymbol{y}$

$$\begin{bmatrix} 126 \\ 874 \\ 126 \\ 126 \\ 126 \\ 874 \\ 874 \\ 254 \\ 356 \\ 356 \\ 356 \\ 264 \\ 264 \\ 874 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0.4 & 0.44 & 0 & 0.27 \\ 0 & 0 & 1 & 0 & 0.6 & 0 & 0.55 & 0.40 \\ 0 & 0 & 1 & 0 & 0.6 & 0 & 0.55 & 0.40 \\ 0 & 0 & 1 & 0 & 0.6 & 0 & 0.55 & 0.40 \\ 0 & 0 & 0 & 1 & 0 & 0.56 & 0.45 & 0.33 \\ 0 & 0 & 0 & 1 & 0 & 0.56 & 0.45 & 0.33 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \bullet \begin{bmatrix} 126 \\ 93 \\ 122 \\ 79 \\ 157 \\ 126 \\ 138 \\ 159 \end{bmatrix}$$

Fig. 5.5.5 Relationship between link, path and hyperpath flows for the network of Fig. 5.5.1

Mixed pre-trip/en-route behavior assignment can therefore be expressed by relations (5.5.7-8) and (5.5.3-4), together with the cost functions (5.2.2). It clearly follows from this formulation that the pre-trip assignment models expressed by relations (5.2.1,3.5.6) are special cases which can be obtained from mixed pre-trip/en route assignment models by setting $\boldsymbol{\Omega} = \boldsymbol{I}$, from which it follows that $\boldsymbol{\Delta} = \boldsymbol{\Lambda}$, $\boldsymbol{y} = \boldsymbol{h}$ and $\boldsymbol{x} = \boldsymbol{g}$. In fact, in pre-trip assignment each hyperpath corresponds to a single path (simple hyperpath), and en-route choices are not considered.

The set of feasible hyperpaths and link flows S_y and S_f are defined, as in Section 5.2.3, by:

$$S_y = \{\boldsymbol{y} : \boldsymbol{y}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T \boldsymbol{y}_{od} = d_{od}\ \forall od\} \tag{5.5.9}$$

$$S_f = \{\boldsymbol{f} = \boldsymbol{\Lambda}\boldsymbol{y},\ \boldsymbol{y} \in S_y\} \tag{5.5.10}$$

The *Uncongested Network assignment* models described in Section 5.3 can therefore easily be extended to the case of mixed pre-trip/en-route choice. In particular, the uncongested network assignment model in terms of link flows can be expressed by an equation similar to (5.3.2):

$$\boldsymbol{f}_{UN} = \boldsymbol{f}_{UN}(\boldsymbol{c}; \boldsymbol{d}) = \sum_{od} d_{od} \boldsymbol{\Lambda}_{od} \boldsymbol{q}_{,od}(-\boldsymbol{\Lambda}_{od}^T \boldsymbol{c} - \boldsymbol{x}_{od}^{NA}) \tag{5.5.11}$$

Uncongested Network assignment models with mixed pre-trip/en-route behavior can be probabilistic or deterministic depending on the hyperpath choice model adopted. In the case of probabilistic path choice behavior, the Stochastic Uncongested Network (*SUN*) assignment models can be expressed by a function similar to (5.3.3):

$$\boldsymbol{f}_{SUN} = \boldsymbol{f}_{SUN}(\boldsymbol{c}; \boldsymbol{d}) = \sum_{od} d_{od} \boldsymbol{\Lambda}_{od} \boldsymbol{q}_{od}(-\boldsymbol{\Lambda}_{od}^T \boldsymbol{c} - \boldsymbol{x}_{od}^{NA}) \qquad \forall \boldsymbol{c} \tag{5.5.12}$$

which retains the properties of continuity and monotonicity discussed in Section 5.4.1.

In the case of deterministic path choice behavior, the relationship between hyperpath flow and costs can be expressed with a system of inequalities similar to (5.3.4):

$$\boldsymbol{x}^T (\boldsymbol{y} - \boldsymbol{y}_{DUN}) \geq 0 \quad \forall \boldsymbol{y} \in S_y \tag{5.5.13}$$

If non-additive path costs are not explicitly considered, by substituting equ. (5.5.7b) and (5.5.8b) in (5.5.13) it follows:

$$\boldsymbol{c}^T (\boldsymbol{f} - \boldsymbol{f}_{DUN}) + (\boldsymbol{x}^{NA})^T (\boldsymbol{y} - \boldsymbol{y}_{DUN}) \geq 0 \qquad \forall \boldsymbol{f} = \boldsymbol{\Lambda}\boldsymbol{y},\ \forall \boldsymbol{y} \in S_y \tag{5.5.14}$$

Because of the presence of non-additive costs, the considerations made in Section 5.3.2 hold and (5.5.14) can be expressed in terms of link flows $\boldsymbol{f}_{DUN}$ and total non-additive cost $X_{DUN} = (\boldsymbol{x}^{NA})^T \boldsymbol{y}_{DUN}$ corresponding to the deterministic assignment:

$$\boldsymbol{c}(\boldsymbol{f}_{DUN})^T(\boldsymbol{f} - \boldsymbol{f}_{DUN}) + (X - X_{DUN}) \geq 0 \quad \forall \boldsymbol{f} = \boldsymbol{\Lambda}\boldsymbol{y}, \forall X = (\boldsymbol{x}^{NA})^T \boldsymbol{y}, \forall \boldsymbol{y} \in S_y$$

Additionally, the *rigid demand equilibrium assignment* models described in Section 5.4 can easily be extended to the case of mixed pre-trip/en-route path choice. It is usually assumed, parallel to the case of non-additive path costs, that the non-additive hyperpath costs, and in particular waiting times at stops (in the case of transit systems), are not affected by congestion; that is, they do not depend on the link flows[21]. Under this hypothesis, by combining the supply model (5.2.4) with the demand model (5.5.5), a system of equations in terms of equilibrium *path variables*, namely costs, $\boldsymbol{g}^*$, and flows, $\boldsymbol{h}^*$, is obtained:

$$\begin{aligned} \boldsymbol{g}^* &= \boldsymbol{\Delta}^T \boldsymbol{c}(\boldsymbol{\Delta}\boldsymbol{h}^*) \\ \boldsymbol{h}^* &= \boldsymbol{\Omega}\boldsymbol{Q}(-\boldsymbol{\Omega}^T\boldsymbol{g}^* - \boldsymbol{x}^{NA})\,\boldsymbol{d} \end{aligned}$$

An analogous formulation in terms of equilibrium *hyperpath variables*, again costs and flows, is also possible:

$$\begin{aligned} \boldsymbol{x}^* &= \boldsymbol{\Lambda}^T \boldsymbol{c}(\boldsymbol{\Lambda}\boldsymbol{y}^*) + \boldsymbol{x}^{NA} \\ \boldsymbol{y}^* &= \boldsymbol{Q}(-\boldsymbol{x}^*)\,\boldsymbol{d} \end{aligned}$$

As in the case of assignment with fully pre-trip path choice behavior, an equivalent formulation in terms of link variables can be expressed by the system of equations obtained by combining the uncongested network assignment map (5.5.11) with the cost functions (5.5.2):

$$\begin{aligned} \boldsymbol{c}^* &= \boldsymbol{c}(\boldsymbol{f}^*) \\ \boldsymbol{f}^* &= \boldsymbol{f}_{UN}(\boldsymbol{c}^*; \boldsymbol{d}) = \boldsymbol{\Lambda}\,\boldsymbol{Q}(-\boldsymbol{\Lambda}^T \boldsymbol{c}^* - \boldsymbol{x}^{NA})\,\boldsymbol{d} \end{aligned}$$

In the case of Stochastic User Equilibrium (SUE) a fixed-point model similar to model (5.4.2) in link flows is obtained:

$$\boldsymbol{f}^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{f}^*); \boldsymbol{d}) = \sum_{od} d_{od}\,\boldsymbol{\Lambda}_{od}\,\boldsymbol{q}_{,od}(-(\boldsymbol{\Lambda}_{od}^T \boldsymbol{c}(\boldsymbol{f}^*) + \boldsymbol{x}_{od}^{NA})) \qquad (5.5.15)$$

with $\boldsymbol{f}^* \in S_f$

Stochastic equilibrium can be also formulated with fixed-point models with path or hyperpath flow variables, or with link, path or hyperpath costs variables; these formulations are not reported here for the sake of brevity. Under the assumption of flow-independent non-additive costs, the conditions of existence and uniqueness analyzed in Section 5.4.1 still hold; in particular, the cost-flow functions for on-

board, access, boarding and alighting links must be respectively continuous and/or strictly increasing[22]. The extension of the results described for the case of flow-dependent non-additive costs (such is waiting costs) is not straightforward and will not be pursued here.

Deterministic User Equilibrium (DUE) assignment can be analyzed with variational inequality models. In particular, denoting the hyperpath cost functions with $\boldsymbol{x}(\boldsymbol{y}) = \boldsymbol{\Lambda}^T \boldsymbol{c}(\boldsymbol{\Lambda}\boldsymbol{y}) + \boldsymbol{x}^{NA}$, models similar to the variational inequality (5.4.3,5) are obtained:

$$\boldsymbol{x}(\boldsymbol{y}^*)^T(\boldsymbol{y} - \boldsymbol{y}^*) \geq 0 \quad \forall \boldsymbol{y} \in S_y \tag{5.5.16}$$

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) + (\boldsymbol{x}^{NA})^T(\boldsymbol{y} - \boldsymbol{y}^*) \geq 0 \quad \forall \boldsymbol{f} = \boldsymbol{\Lambda}\boldsymbol{y}, \forall \boldsymbol{y} \in S_y \tag{5.5.17}$$

Non-additive hyperpath costs can be handled as described in Section 5.4.2. The above expression can be formulated in terms of link flows $\boldsymbol{f}^*$ and total non-additive cost X^*:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f} - \boldsymbol{f}^*) + (X - X^*) \geq 0 \quad \forall \boldsymbol{f} = \boldsymbol{\Lambda}\boldsymbol{y}, \forall X = (\boldsymbol{x}^{NA})^T\boldsymbol{y}, \forall \boldsymbol{y} \in S_y$$

The optimization models described in the previous sections and in the appendix for deterministic or stochastic assignment can also be easily applied in this case, within the limits of the assumptions.

5.6. Elastic demand User Equilibrium assignment models*

Elastic demand assignment models assume that demand flows depend on transportation costs. These models simulate supply-demand interactions when user choice behavior other than path (such as mode, destination, etc.) is influenced by path cost variations due to variations in congested link costs[23]. The dependence of demand on cost $\boldsymbol{c}$ is expressed by the demand models described in Chapter 4.

If demand models are based on random utility theory, the demand flow for each O-D pair generally depends on the values of the (systematic) utilities associated with the paths available for the various O-D pairs, through the EMPU of path choice. This can be seen as an "average" over the systematic utilities (i.e. costs) of the available paths. This is described in section 3.5 and in section 4.2 on the general structure of demand models.

For uncongested networks, elastic demand assignment is not meaningful, since path costs, EMPUs and thus the demand flows are independent of link flows. Link and path flows can then be obtained through the Uncongested Network assignment models described in Section 5.3. On the other hand, for congested networks, costs depend on flows, and a further mutual dependence between flows and costs is introduced through the demand function as shown in Fig. 5.6.1 which distinguishes between internal and external approaches.

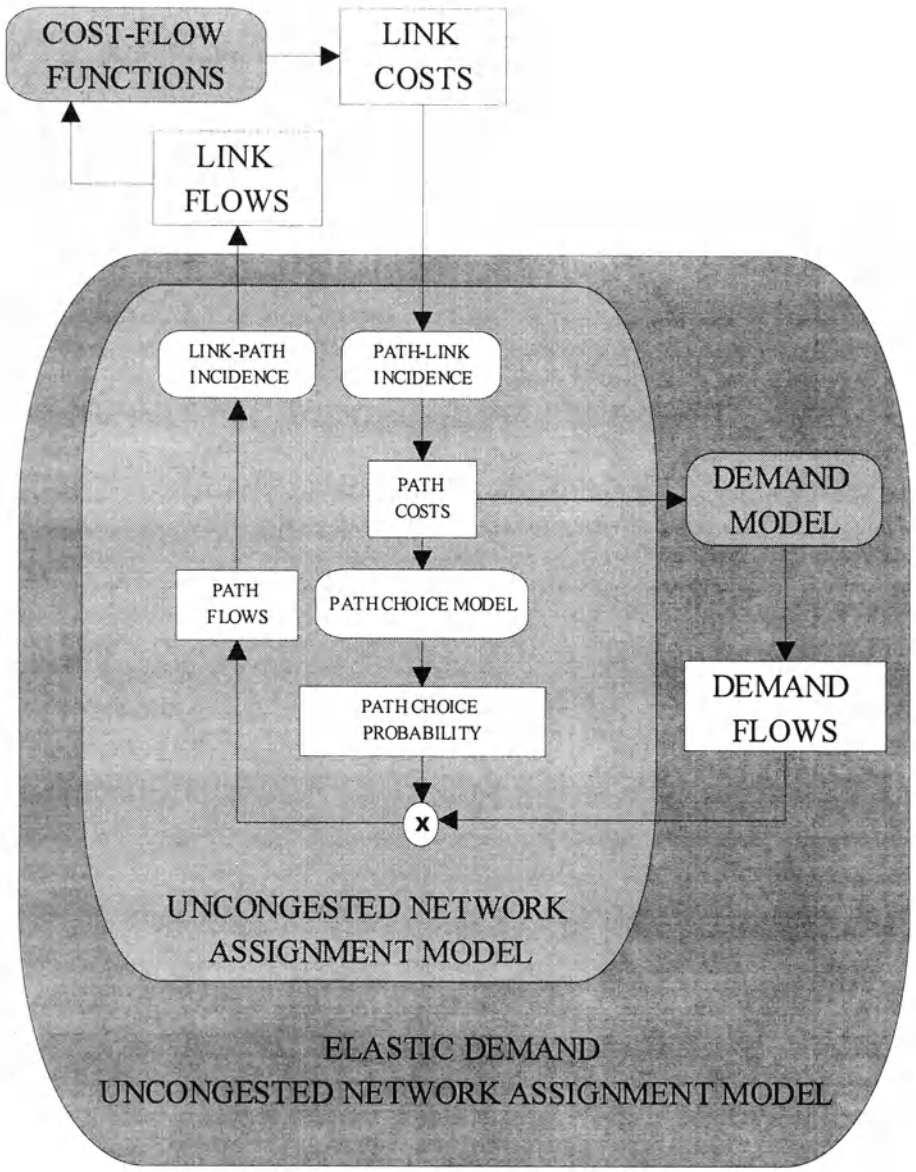


Fig. 5.6.1.a Schematic representation of elastic demand equilibrium assignment models (internal approach).

For elastic demand equilibrium assignment, it is useful to distinguish between two cases. In *single-mode assignment*, dealt with in section 5.6.1, there is one mode for which link costs depend on flows, and either the demand elasticity does not depend on modal split at all or link costs for all other modes are not congestion-dependent. In the second case, level of service attributes of uncongested modes are known before the solution of the assignment model and play a role similar to demand model parameters for the congested mode. Once the congested mode equilibrium assignment has been solved and the cost attributes of this mode have been determined, the demand for the other (uncongested) modes can be obtained and assigned through uncongested network assignment models, one mode at a time.

In *multi-modal assignment*, dealt with in Section 5.6.2, there is more than one mode with link costs depending on flows (congested modes). In this case, the cost attributes of congested modes cannot be known before the solution of the assignment model, and it is necessary to solve the equilibrium assignment problem simultaneously (at least for congested modes). Note that congested modes may have separate supply (network) and path choice models.

To clarify the difference between the two types of elastic demand assignment, consider the case of choice between two modes, car and bus. If bus travel times are independent of the relative link flows, the level-of service attributes of this mode are independent of congestion. They can be calculated through the network model and then used as attributes of the mode choice model providing demand flows for the car mode. The known costs of the bus mode and the cost functions for the car mode allow the specification of a single-mode assignment on congested network with elastic demand for the car mode. When this model is solved and the car mode equilibrium attributes are found, the bus mode demand flows will be determined and an uncongested network assignment can be performed. Vice versa, if the costs of both modes depend on the flows, it is necessary to assign the demand of both modes at the same time in order to find the congested cost pattern for each of them. These costs have to be consistent with mode choice, path choices and the network flows of the two modes.

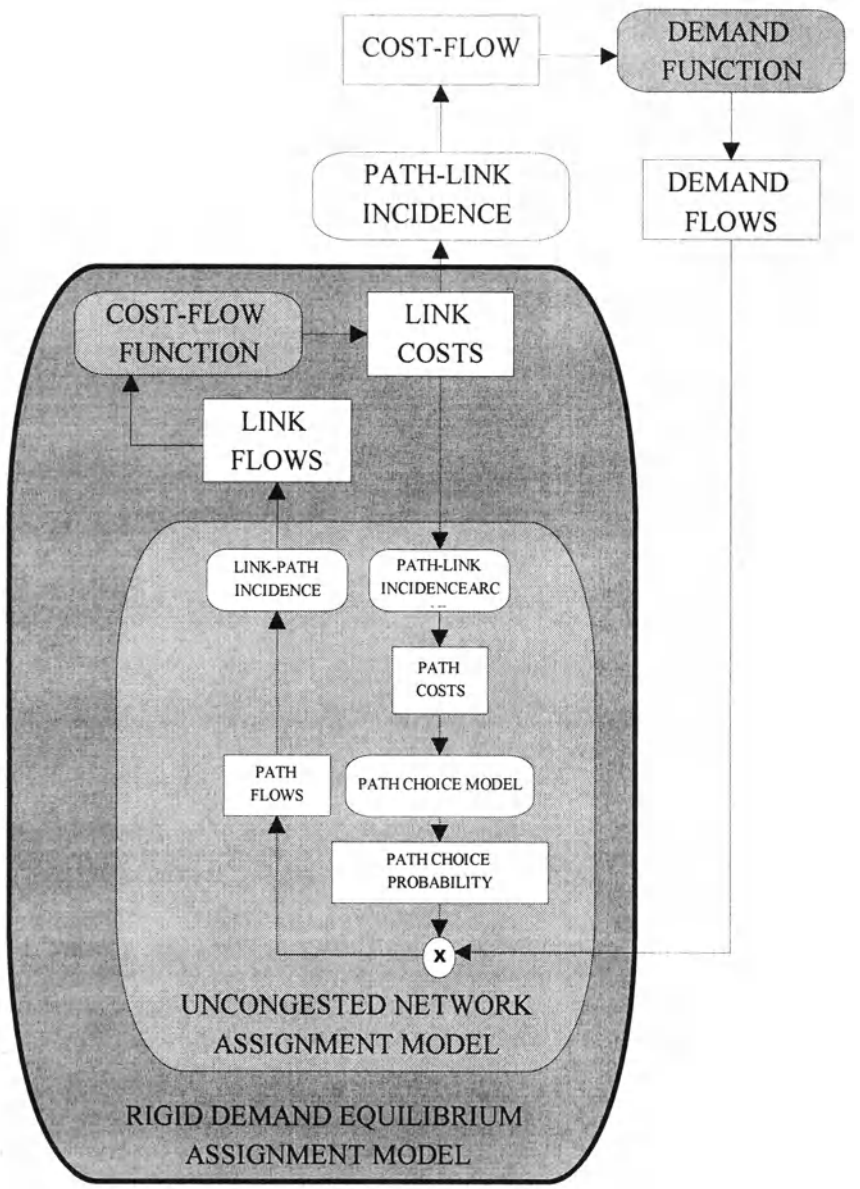


Fig. 5.6.1b Schematic representation of elastic demand equilibrium assignment models (external approach).

5.6.1. Single-mode assignment models

With reference to a single mode m and to the time band h (not explicitly indicated in the following) the *demand function* for the *od* pair can be expressed as:

$$d_{od} = d_{od}[mh] = d_{od}(\boldsymbol{s}) \quad \forall od$$

or in matrix terms:

$$\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{s})$$

where:

$\boldsymbol{d}$ is the demand flow vector with an element d_{od} for each *od* pair;
$\boldsymbol{s}$ is the path choice EMPU vector (for the mode m in the time band h) with an element s_{od} for each *od* pair.

The demand function simulates the dependence between demand flows and EMPU in a general sense, and will vary depending on the choice dimensions considered elastic with respect to congestion costs. For example, if demand elasticity is relative to destination choice, the demand flow d_{od} depends only on the elements of the vector $\boldsymbol{s}$ for the O-D pairs with origin in the zone o, $d_{od} = d_{od}(s_{od1},..., s_{odn},...)$. If the demand flow, d_{od}, of the *od* pair depends only on the EMPU, s_{od}, of the same *od* pair, we have the special case of separable demand functions, $d_{od} = d_{od}(s_{od})$; this may arise in the case of elastic trip frequency, or emission demand models.

The EMPU depends in turn on the values of the path systematic utility through the relation (5.2.8) given in Section 5.2:

$$s_{od} = s_{od}(V_{od}) \quad \forall od$$

Note that the EMPU is defined and measured homogeneously with the utility. So it is negative in the case of a path choice model since the systematic utility of each path is generally negative, being the opposite to the corresponding systematic cost. From the systematic utility expression (5.2.5) it follows:

$$d_{od} = d_{od}(\boldsymbol{s}\,(\boldsymbol{V} = -\boldsymbol{g})) \quad \forall od$$

or in matrix notation:

$$\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{s}(\boldsymbol{V} = -\boldsymbol{g})) \tag{5.6.1}$$

If destination choice, for example, is simulated with a Logit model (parameter θ_1) and path choice is simulated with a Logit model (parameter θ_2), an elementary specification of the previous expression could be:

$$d_{od} = d_{o.}\ exp((\beta_1 A_d + \beta_2 s_{od}) / \theta_1) / \Sigma_j\ exp((\beta_1 A_j + \beta_2 s_{oj}) / \theta_1) \quad \forall od$$

$$s_{od} = \theta_2 \, ln(\Sigma_{k \in Kod} \, exp(-g_k/\theta_2)) \qquad \forall od$$

where:

d_o, is the total flow leaving from zone o, assumed constant;
A_d is the attraction attribute of the destination zone d;
β_1, are the homogenization coefficients in the systematic utility V_{od} between the
β_2 zones o and d.

In elastic demand assignment models, described below, it is assumed that the demand flow, d_{od}, for each od pair is non-negative and bounded above by a positive value, that is $d_{od} \in [0, d_{od,max}]$, and it is therefore possible to define the set of feasible demand flow vectors as:

$$S_d = \{\boldsymbol{d} : d_{od} \in [0, d_{od,max}] \; \forall od \}$$

Under this hypothesis, the set of feasible path flows S_h, and of feasible link flows S_f, described in Section 5.2 are compact and convex (and non-empty if the network is connected) as in the case of rigid demand.

The demand model in the case of elastic demand becomes:

$$\boldsymbol{h}_{od} = d_{od}(\boldsymbol{s}(-\boldsymbol{g})) \, \boldsymbol{p}_{od}(-\boldsymbol{g}_{od}) \quad \forall od \qquad (5.6.2a)$$

$$\boldsymbol{h} = \boldsymbol{P}(-\boldsymbol{g}) \, \boldsymbol{d}(\boldsymbol{s}(-\boldsymbol{g})) \qquad (5.6.2b)$$

Note that expression 5.6.2 is the equivalent of 5.2.7 in the case of rigid demand. On the other hand, the supply model remains unchanged as expressed by the relation 5.2.4.

The (single-mode) equilibrium approach in the case of elastic demand assumes that the state of the system can be represented by a path flow configuration $\boldsymbol{h}^*$ which is mutually consistent with the corresponding path costs $\boldsymbol{g}^*$, as defined by the supply model (5.2.4) and by the demand model (5.6.2):

$$\boldsymbol{g}^* = \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^*) + \boldsymbol{g}^{NA}$$
$$\boldsymbol{h}^* = \boldsymbol{P}(-\boldsymbol{g}^*) \, \boldsymbol{d}(\boldsymbol{s}(-\boldsymbol{g}^*))$$

The corresponding equilibrium demand flows $\boldsymbol{d}^*$ are given by (5.6.1). An equivalent formulation of the elastic demand single-mode equilibrium assignment model can be expressed in terms of link variables. In this case, the system of equations in terms of equilibrium link flows, $\boldsymbol{f}^*$, is obtained by combining the cost functions (5.2.2) with the equation obtained by combining the uncongested network assignment map (5.3.2), the demand function (5.6.1), and the path costs expression (5.2.1):

$$c^* = c(f^*)$$
$$f^* = f_{UN\text{-}EL}(c^*; d) = \Delta P(-\Delta^T c^* - g^{NA})\, d(s(-\Delta^T c^* - g^{NA}))$$

The circular dependence between demand flows and costs can also be expressed externally to the equilibrium between (link and path) flows and costs. At the inner level, for a given vector of demand flows, (rigid demand) equilibrium link flows and costs are defined by the path choice model and by the cost functions. At the outer level, the equilibrium between the costs resulting from the (rigid demand) equilibrium assignment and the demand flows defined by the demand functions is defined. Let

$f_{UE\text{-}RIG} = f_{UE\text{-}RIG}(d)$ be the implicit correspondence between the rigid demand equilibrium link flows, $f_{UE\text{-}RIG}$, and the demand flows d. This correspondence is defined by the solution of one of the models described in Section 5.4. It is a function (on-to-one correspondence) if equilibrium flows are unique.

External elastic demand equilibrium assignment can therefore be formulated with a system of non-linear equations:

$$d^* = d(s(-\Delta^T c(f^*)))$$
$$f^* = f_{UE\text{-}RIG}(d^*)$$

Combining the two previous equations, a fixed-point problem (with an implicitly defined function) is obtained with respect to the demand flows, d^* or link flows, f^*. Formulations with respect to link cost or EMPU are also possible. The external approach can be adopted to define solution procedures, as will be seen in Chapter 7, but it is difficult to analyze theoretically.

The analysis of elastic demand equilibrium assignment can easily be carried out for the internal approach through direct extension of the rigid demand equilibrium assignment models described in section 5.4[24], distinguishing the cases of stochastic and deterministic equilibrium.

5.6.1.1. Elastic demand single-mode Stochastic User Equilibrium models

The fixed-point model with respect to path flows (5.4.1) for rigid demand stochastic equilibrium can easily be extended to the case of elastic demand by combining the supply model (5.2.4) and the demand model (5.6.2):

$$h^* = P(-\Delta^T c(\Delta h^*) - g^{NA})\; d(s(-\Delta^T c(\Delta h^*) - g^{NA})) \qquad (5.6.3)$$

with $h^* \in S_h$

Also, the equivalent fixed-point model with respect to link flows (5.4.2) for the rigid demand SUE can easily be extended to elastic demand:

$$\boldsymbol{f}^* = \Sigma_{od}\, d_{od}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}(\boldsymbol{f}^*) - \boldsymbol{g}^{NA}))\, \Delta_{od}\, \boldsymbol{p}_{od}(-\Delta_{od}{}^T \boldsymbol{c}(\boldsymbol{f}^*) - \boldsymbol{g}_{od}{}^{NA}) \qquad (5.6.4)$$

with $\boldsymbol{f}^* \in S_f$

Equilibrium link costs are given by $\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$, and therefore the corresponding demand flows are given by $d_{od}^* = d_{od}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}^* - \boldsymbol{g}^{NA}))$.

The analysis of the existence and uniqueness of the solutions is a straightforward extension of the results given in Section 5.4.1. It requires explicit assumptions on the demand functions sufficient to ensure continuity and monotonicity of the stochastic uncongested network assignment function (with elastic demand) given by:

$$\boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c}) = \boldsymbol{f}_{SUN}(\boldsymbol{c}^*;\, \boldsymbol{d} = d(\boldsymbol{s}(-\Delta^T \boldsymbol{c} - \boldsymbol{g}^{NA}))) =$$
$$= \Sigma_{od}\, d_{od}(\boldsymbol{s}(-\Delta^T \boldsymbol{c} - \boldsymbol{g}^{NA}))\, \Delta_{od}\, \boldsymbol{p}_{od}(-\Delta_{od}{}^T \boldsymbol{c} - \boldsymbol{g}_{od}{}^{NA})$$

In what follows, existence and uniqueness are analyzed explicitly only for equilibrium link flows. These conditions also ensure the existence and uniqueness of the corresponding equilibrium link costs, $\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*)$, path costs and flows, $\boldsymbol{g}^*$ and $\boldsymbol{h}^*$, obtained with the expressions reported in Section 5.2, as well as demand flows, d_{od}^*.

Existence of the elastic demand stochastic user equilibrium. The fixed-point model with respect to link flows (5.6.4) has at least one solution if path choice probability functions, $\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od})$, EMPU functions, $s_{od} = s_{od}(\boldsymbol{V}_{od})$, and demand functions, $d_{od} = d_{od}(\boldsymbol{s})$, composing the SUN function, $\boldsymbol{f} = \boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c})$, and cost functions, $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$, are continuous (also assuming that each O-D pair is connected and demand flows are limited).

The proof is similar to that in Section 5.4.1. for rigid demand.

Monotonicity of the elastic demand stochastic uncongested network function. If path choice models are defined by non-decreasing monotonic functions with respect to the systematic utilities, as is the case with probabilistic additive models (with $|\Sigma| \neq 0$, see section 3.5), and demand functions are non-negative, upper bounded and non-decreasing with respect to the EMPU:

$$[\boldsymbol{d}(\boldsymbol{s}') - \boldsymbol{d}(\boldsymbol{s}'')]^T (\boldsymbol{s}' - \boldsymbol{s}'') \geq 0 \qquad \forall \boldsymbol{s}', \boldsymbol{s}''$$

the elastic demand stochastic uncongested network assignment function is non-increasing monotone with respect to link costs. Thus if the cost of one (or more) link increases, the flow (or flows) of that (or those) link does not increase. This property is expressed formally as:

$$(\boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c}') - \boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c}''))^T (\boldsymbol{c}' - \boldsymbol{c}'') \leq \tilde{0} \quad \forall \boldsymbol{c}', \boldsymbol{c}''$$

Under the assumptions made, given the two systematic utility vectors, $\boldsymbol{V}_{od}'$

and V_{od}'', corresponding to the paths which connect the pair od, the following relations involving the corresponding path choice probabilities and the EMPU hold (see section 3.5):

$$\boldsymbol{p}_{od}(V_{od}')^T (V_{od}' - V_{od}'') \geq s_{od}(V_{od}') - s_{od}(V_{od}'')$$
$$s_{od}(V_{od}') - s_{od}(V_{od}'') \geq \boldsymbol{p}_{od}(V_{od}'')^T (V_{od}' - V_{od}'')$$

Letting $s_{od}' = s_{od}(V_{od}')$ and $s_{od}'' = s_{od}(V_{od}'')$, multiplying the first relation by $d_{od}(\boldsymbol{s}') \geq 0$ and the second by $d_{od}(\boldsymbol{s}'') \geq 0$ gives:

$$d_{od}(\boldsymbol{s}')\, \boldsymbol{p}_{od}(V_{od}')^T (V_{od}' - V_{od}'') \geq d_{od}(\boldsymbol{s}') (s_{od}(V_{od}') - s_{od}(V_{od}''))$$
$$d_{od}(\boldsymbol{s}'') (s_{od}(V_{od}') - s_{od}(V_{od}'')) \geq d_{od}(\boldsymbol{s}'')\, \boldsymbol{p}_{od}(V_{od}'')^T (V_{od}' - V_{od}'')$$

from which, summing over all the O-D pairs, it follows that:

$$\Sigma_{od}\, d_{od}(\boldsymbol{s}')\, \boldsymbol{p}_{od}(V_{od}')^T (V_{od}' - V_{od}'') \geq \Sigma_{od}\, d_{od}(\boldsymbol{s}') (s_{od}(V_{od}') - s_{od}(V_{od}''))$$
$$\Sigma_{od}\, d_{od}(\boldsymbol{s}'') (s_{od}(V_{od}') - s_{od}(V_{od}'')) \geq \Sigma_{od}\, d_{od}(\boldsymbol{s}'')\, \boldsymbol{p}_{od}(V_{od}'')^T (V_{od}' - V_{od}'')$$

Furthermore, for the monotonicity of the demand functions, it follows that:

$$\Sigma_{od}\, d_{od}(\boldsymbol{s}') (s_{od}(V_{od}') - s_{od}(V_{od}'')) \geq \Sigma_{od}\, d_{od}(\boldsymbol{s}'') (s_{od}(V_{od}') - s_{od}(V_{od}''))$$

Therefore, the following is expression obtained:

$$\Sigma_{od}\, d_{od}(\boldsymbol{s}')\, \boldsymbol{p}_{od}(V_{od}')^T (V_{od}' - V_{od}'') \geq \Sigma_{od}\, d_{od}(\boldsymbol{s}'')\, \boldsymbol{p}_{od}(V_{od}'')^T (V_{od}' - V_{od}'')$$

from which, letting $\boldsymbol{h}_{od}' = d_{od}(\boldsymbol{s}')\boldsymbol{p}_{od}'$ and $\boldsymbol{h}_{od}'' = d_{od}(\boldsymbol{s}'')\boldsymbol{p}_{od}''$, it is deduced:

$$\Sigma_{od} (\boldsymbol{h}_{od}' - \boldsymbol{h}_{od}'')^T (V_{od}' - V_{od}'') \geq 0$$

Given two different link cost vectors, $\boldsymbol{c}'$ and $\boldsymbol{c}''$, let

$$\boldsymbol{g}_{od}' = \Delta_{od}^T \boldsymbol{c}' + \boldsymbol{g}_{od}^{NA} \qquad V_{od}' = -\boldsymbol{g}_{od}'$$
$$\boldsymbol{g}_{od}'' = \Delta_{od}^T \boldsymbol{c}'' + \boldsymbol{g}_{od}^{NA} \qquad V_{od}'' = -\boldsymbol{g}_{od}''$$

Therefore, parallel to the exposition in Section 5.3.1, with

$$\boldsymbol{f}' = \Sigma_{od}\, \Delta_{od} \boldsymbol{h}_{od}' \qquad \boldsymbol{f}'' = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}''$$

it is finally obtained: $(\boldsymbol{f}' - \boldsymbol{f}'')^T (\boldsymbol{c}' - \boldsymbol{c}'') \leq 0$

Note that, under these assumptions, the Jacobian $\boldsymbol{Jac}[\boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c})]$ is symmetric negative semi-definite since the Jacobian $\boldsymbol{Jac}[\boldsymbol{p}_{od}(\boldsymbol{V}_{od})]$ is symmetric positive semi-definite (see section 3.5).

Uniqueness of elastic demand stochastic user equilibrium. The fixed-point model in terms of link flows (5.6.4) has at the most one solution if link cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are strictly increasing with respect to the feasible link flows:

$$[\boldsymbol{c}(\boldsymbol{f}') - \boldsymbol{c}(\boldsymbol{f}'')]^T (\boldsymbol{f}' - \boldsymbol{f}'') > 0 \quad \forall \boldsymbol{f}' \neq \boldsymbol{f}'' \in S_f$$

If demand functions are non-negative, bounded above, and non-decreasing with respect to the EMPU:

$$[\boldsymbol{d}(\boldsymbol{s}') - \boldsymbol{d}(\boldsymbol{s}'')]^T (\boldsymbol{s}' - \boldsymbol{s}'') \geq 0 \quad \forall \boldsymbol{s}', \boldsymbol{s}''$$

and if path choice models are additive, in the sense defined in section 5.3.1, and expressed by continuous functions $\boldsymbol{p}_{od} = \boldsymbol{p}_{od}(\boldsymbol{V}_{od})$ with continuous first partial derivatives .

The proof is similar to the one provided for rigid demand in section 5.4.1. In fact, under the above assumptions, the elastic demand SUN function, $\boldsymbol{f}_{SUN\text{-}EL}(\boldsymbol{c})$, is non-increasing monotone with respect to the link costs.

The considerations made in section 5.4.1 on the existence and uniqueness of the solutions, and on the continuity and monotonicity of the cost functions, can be extended directly to elastic demand models. As for the demand functions, their monotonicity implies that variations in path cost induce opposite variations in EMPU's and therefore in demand flows. In other words, the increase of a link cost, and therefore of the cost of the paths including it, cannot induce an increase in the demand flows between the O-D pairs connected by these paths. This property is always guaranteed if the demand functions are defined through probabilistic choice models, such as trip distribution models, which are additive with respect to the EMPU of path choice (also deterministic demand models satisfy the monotonicity requirement).

5.6.1.2. Elastic demand single-mode Deterministic User Equilibrium models

When path choice behavior is simulated with a deterministic model, the EMPU is given, as stated in section 3.5, by the maximum systematic utility, or the opposite of the minimum path cost:

$$s_{od} = s_{od}(\boldsymbol{V}_{od}) = \max_{k \in Kod} (V_{od,k}) = -\min_{k \in Kod} (g_{od,k}) = -Z_{od} \quad \forall od$$

where:

$Z_{od} = -s_{od}$ is the minimum cost of the paths connecting the pair od;
$\boldsymbol{Z} = -\boldsymbol{s}$ is the vector of the minimum path costs between all the O-D pairs, with an element for each O-D pair.

The demand functions are, in the case of deterministic assignment, usually expressed in terms of minimum cost i.e. the opposite of the EMPU (still using notation $d(.)$ for demand function):

$$d_{od} = d_{od}(-\boldsymbol{s}) = d_{od}(\boldsymbol{Z}) \quad \forall \boldsymbol{Z} \quad \forall od$$

or equivalently

$$\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{Z}) \quad \forall \boldsymbol{Z} \tag{5.6.5a}$$

As an example, consider the case of a Logit model simulating destination choice analogous to the one described previously, while path choice is simulated with a deterministic model. Expression (5.6.5a) becomes:

$$d_{od} = d_o \, exp((\beta_1 A_d - \beta_2 Z_{od})/\theta_1) \,/\, \Sigma_j \, exp((\beta_1 A_j - \beta_2 Z_{oj})/\theta_1) \quad \forall od$$

where:

d_o is the total flow leaving zone o, assumed constant;
A_d is the attraction attribute of the destination zone d;
θ_l is the Logit parameter;
β_1, β_2 are the homogenization coefficients in the systematic utility between zones o and d.

The indirect formulation of rigid demand deterministic equilibrium through variational inequality models, described in paragraph 5.4.2, can be extended to elastic demand. For this purpose, it is necessary to assume that the demand functions (5.6.5) are invertible[25], i.e. that it is possible to define the (vector) *inverse demand function*[26] giving for each demand flow vector, $\boldsymbol{d}$, the corresponding vector of minimum path costs $\boldsymbol{Z}$. This is the vector of minimum path costs vector that, through the demand function, generates the demand vector, $\boldsymbol{d}$,:

$$\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{d}) \quad \forall \boldsymbol{d} \in S_d \tag{5.6.5b}$$

The inverse demand function (5.6.5b) has the same properties of continuity and monotonicity as the demand function (5.6.5a). In particular, it is strictly decreasing if (and only if) the demand function (5.6.5a) is strictly decreasing. Thus for an increase of demand flows, the inverse demand function associates a decrease in costs. This property is guaranteed if the demand function is defined by additive probabilistic choice models with respect to mininum costs (or by deterministic models).

The variational inequality formulation of elastic demand deterministic equilibrium assignment can be achieved by extending the path flows model (5.4.3) described in section 5.4.2 for rigid demand. In the case of elastic demand, this model becomes (excluding non-additive path costs for simplicity of notation):

$$\boldsymbol{g}(\boldsymbol{h}^*)^T(\boldsymbol{h}-\boldsymbol{h}^*) - \boldsymbol{Z}(\boldsymbol{d}^*)^T(\boldsymbol{d}-\boldsymbol{d}^*) \geq 0 \quad \forall \boldsymbol{h} \in S_h \quad \forall \boldsymbol{d} \in S_d \tag{5.6.6}$$

In fact, condition (3.5.11a) on deterministic choice probabilities, $\boldsymbol{p}_{DET,od}$, (as introduced in section 3.5) applied to each *od* pair yields:

$$\boldsymbol{V}_{od}^T \boldsymbol{p}_{DET,od} = max(\boldsymbol{V}_{od}) \qquad \forall od$$

Given path costs $\boldsymbol{g}_{od}^*$, let $Z_{od}^* = min\ (\boldsymbol{g}_{od}^*)$ be the minimum path cost for each *od* pair. Assuming $\boldsymbol{V}_{od}^* = -\boldsymbol{g}_{od}^*$ yields $max(\boldsymbol{V}_{od}^*) = -Z_{od}^*$. Furthermore, let d_{od}^* be the demand flow corresponding to minimum cost Z_{od}^*, that is $\boldsymbol{Z}^* = \boldsymbol{Z}(\boldsymbol{d}^*)$ to be consistent with inverse demand function. Multiplying the above equation by the non-negative demand flow $d_{od}^* \geq 0 \quad \forall od$ yields:

$$(\boldsymbol{g}_{od}^*)^T \boldsymbol{h}_{DET,od} = Z_{od}^* d_{od}^* \qquad \forall od \tag{a}$$

since $\boldsymbol{h}_{DET,od} = d_{od}^* \boldsymbol{p}_{DET,od} \quad \forall od.$

Generally, the following condition also holds (see section 3.5):

$$\boldsymbol{V}_{od}^T \boldsymbol{p}_{od} \leq max(\boldsymbol{V}_{od}) \quad \forall \boldsymbol{p}_{od} : \boldsymbol{p}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T\boldsymbol{p}_{od} = 1 \quad \forall od$$

Given the path costs $\boldsymbol{g}_{od}^*$, with $\boldsymbol{V}_{od}^* = -\boldsymbol{g}_{od}^*$ and $max(\boldsymbol{V}_{od}^*) = -Z_{od}^*$, multiplying the above equation by any feasible demand flow, $d_{od} \geq 0 \forall od$, yields:

$$(\boldsymbol{g}_{od}^*)^T \boldsymbol{h}_{od} \leq Z_{od}^* d_{od} \quad \forall \boldsymbol{h}_{od} : \boldsymbol{h}_{od} \geq \boldsymbol{0},\ \boldsymbol{1}^T\boldsymbol{h}_{od} = d_{od} \quad \forall d_{od} \geq 0 \quad \forall od$$

thus $$(\boldsymbol{g}_{od}^*)^T \boldsymbol{g}_{od} \leq Z_{od}^* d_{od} \quad \forall \boldsymbol{h}_{od} : \boldsymbol{h} \in S_h, \forall d_{od} : \boldsymbol{d} \in S_d \quad \forall od \tag{b}$$

since $\boldsymbol{h}_{od} = d_{od}\, \boldsymbol{p}_{od} \ \forall od.$

Subtracting equation (a) from (b) yields:

$$(\boldsymbol{g}_{od}^*)^T (\boldsymbol{g}_{od} - \boldsymbol{g}_{DET,od}) \leq Z_{od}^* (d_{od} - d_{od}^*) \quad \forall \boldsymbol{h}_{od} : \boldsymbol{h} \in S_h, \forall d_{od} : \boldsymbol{d} \in S_d \quad \forall od$$

Summing up the above equation for all the *od* pair with $\boldsymbol{Z}^* = \boldsymbol{Z}(\boldsymbol{d}^*)$, a deterministic demand model with elastic demand is obtained:

$$(\boldsymbol{g}^*)^T (\boldsymbol{h} - \boldsymbol{h}_{DET}) \leq \boldsymbol{Z}(\boldsymbol{d}^*)^T (\boldsymbol{d} - \boldsymbol{d}^*) \quad \forall \boldsymbol{h} \in S_h, \forall \boldsymbol{d} \in S_d$$

Combining the above demand model (d) with the supply model (5.2.4), say $\boldsymbol{g}(\boldsymbol{h}^*) = \boldsymbol{\Delta}^T \boldsymbol{c}(\boldsymbol{\Delta h}^*) + \boldsymbol{g}^{NA}$ relation (5.6.6) is obtained..

Expression (5.6.6) can easily be reformulated in terms of link flows, extending the model (5.4.4) described in section 5.4.2. Expressing equilibrium path costs in terms of link costs according to the supply model, it follows parallel to (5.2.4) that:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T (\boldsymbol{f} - \boldsymbol{f}^*) - \boldsymbol{Z}(\boldsymbol{d}^*)^T (\boldsymbol{d} - \boldsymbol{d}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f \ \forall \boldsymbol{d} \in S_d \tag{5.6.7}$$

The existence of (link or path) flows and costs and the uniqueness of link flows and costs as well as of the demand flows for elastic demand deterministic user equilibrium are guaranteed respectively by the continuity and monotonicity of the cost functions and of the (inverse) demand functions[27].

Existence of elastic demand deterministic user equilibrium. Variational inequalities (5.5.6,7) have at least one solution if the cost functions, defined over the non-empty, compact and convex set of the feasible path or link, flows, and the inverse demand functions, defined over the non-empty, closed and limited interval of demand values, are continuous.

The proof is similar to that described for rigid demand in Section 5.4.1.

Uniqueness of elastic demand deterministic user equilibrium link flows The variational inequality (5.6.7) expressed in terms of link flows has at the most one solution if the link cost functions, $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$, are strictly increasing with respect to the link flows:

$$[\boldsymbol{c}(\boldsymbol{f}') - \boldsymbol{c}(\boldsymbol{f}'')]^T (\boldsymbol{f}' - \boldsymbol{f}'') > 0 \quad \forall \boldsymbol{f}' \neq \boldsymbol{f}'' \in S_f$$

and the inverse demand functions, $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{d})$, are strictly decreasing[28] with respect to the demand flows (i.e. the demand functions are strictly decreasing with respect to the minimum cost):

$$[\boldsymbol{Z}(\boldsymbol{d}') - \boldsymbol{Z}(\boldsymbol{d}'')] (\boldsymbol{d}' - \boldsymbol{d}'') < 0 \quad \forall \boldsymbol{d}' \neq \boldsymbol{d}'' \in S_d$$

The proof, parallel to that described for rigid demand in section 5.4.2, is performed by a *reductio ad absurdum*. If there existed two different equilibrium link flow vector $\boldsymbol{f}_1^* \neq \boldsymbol{f}_2^* \in S_f$ corresponding to two feasible demand flow vectors $\boldsymbol{d}_1^*, \boldsymbol{d}_2^* \in S_d$ (not necessarily different), they both would satisfy (5.6.7) and therefore, with $\boldsymbol{f} = \boldsymbol{f}_2^* \in S_f$ e $\boldsymbol{d} = \boldsymbol{d}_2^* \in S_d$, it would follow:

$$c(\boldsymbol{f}_1^*)^T(\boldsymbol{f}_2^* - \boldsymbol{f}_1^*) - \boldsymbol{Z}(\boldsymbol{d}_1^*)^T(\boldsymbol{d}_2^* - \boldsymbol{d}_1^*) \geq 0$$

Furthermore, also $\boldsymbol{f}_2^*$ e $\boldsymbol{d}_2^*$ would respect (5.6.7) and therefore, with $\boldsymbol{f} = \boldsymbol{f}_1^* \in S_f$ e $\boldsymbol{d} = \boldsymbol{d}_1^* \in S_d$, we would have:

$$c(\boldsymbol{f}_2^*)^T(\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) - \boldsymbol{Z}(\boldsymbol{d}_2^*)^T(\boldsymbol{d}_1^* - \boldsymbol{d}_2^*) \geq 0$$

Summing of the two above relations gives:

$$c(\boldsymbol{f}_1^*)^T(\boldsymbol{f}_2^* - \boldsymbol{f}_1^*) - \boldsymbol{Z}(\boldsymbol{d}_1^*)^T(\boldsymbol{d}_2^* - \boldsymbol{d}_1^*) + c(\boldsymbol{f}_2^*)^T(\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) - \boldsymbol{Z}(\boldsymbol{d}_2^*)^T(\boldsymbol{d}_1^* - \boldsymbol{d}_2^*) \geq 0$$

or

$$[c(\boldsymbol{f}_1^*) - c(\boldsymbol{f}_2^*)]^T(\boldsymbol{f}_1^* - \boldsymbol{f}_2^*) - [\boldsymbol{Z}(\boldsymbol{d}_1^*) - \boldsymbol{Z}(\boldsymbol{d}_2^*)]^T(\boldsymbol{d}_1^* - \boldsymbol{d}_2^*) \leq 0$$

which contradicts the assumption of the monotonicity of the cost functions, and of the inverse demand functions, if $\boldsymbol{d}_1^* \neq \boldsymbol{d}_2^*$. Analogously, if two different vectors of feasible demand flows existed $\boldsymbol{d}_1^* \neq \boldsymbol{d}_2^* \in S_d$, to which corresponded two vectors of equilibrium link flows $\boldsymbol{f}_1^*, \boldsymbol{f}_2^* \in S_f$, this would again result in a contradiction.

Note that, as in the case of rigid demand, the uniqueness of link flows and equilibrium demand does not imply the uniqueness of equilibrium path flows.

Formulation with optimization models. Elastic demand deterministic user equilibrium can also be formulated with optimization models. These allow simple solution algorithms to be used (see Chapter 7). Equivalent optimization models require that cost functions and inverse demand functions have symmetric Jacobians. In particular, assuming for the sake of simplicity the absence of non-additive path costs, the model (5.4.6) can be extended in the following form:

$$(\boldsymbol{f}^*, \boldsymbol{d}^*) = \underset{\substack{\boldsymbol{f} \in S_f \\ \boldsymbol{d} \in S_d}}{argmin} \int_0^f c(\boldsymbol{x})^T d\boldsymbol{x} - \int_0^d \boldsymbol{Z}(\boldsymbol{y})^T d\boldsymbol{y} \qquad (5.6.8)$$

In general, formulation (5.6.8) is of limited use in practice since it is difficult to express the inverse demand function, $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{d})$, in closed form, and therefore to prove the symmetry of its Jacobian.

This condition holds, however, if the demand model is of the Logit type, like the one described at the beginning of this section. In this case, the following holds:

$$\int_0^d \boldsymbol{Z}(y)\, dy = (\theta_1/\beta_2)\, \Sigma_{od}\, (d_{od}\, ln\, d_{od} - d_{od}) + (\beta_1/\beta_2)\, \Sigma_{od}\, (A_d\, d_{od}) \qquad (5.6.9)$$

with $$\Sigma_{od}\, d_{od} = d_o \quad \forall o$$

Analogously, the integral (5.6.9) can explicitly be computed for Logit mode choice models demand with attributes independent of congestion for the other transportation modes.

5.6.2. Multi-mode assignment models

The previous models can be extended to multi-modal assignment in which at least mode choice depends on congested costs for more than one mode. Obviously, in addition to mode and path choice, demand models can be elastic with respect to other choice dimensions, such as frequency and destination. To specify these models it is useful to modify the notation used in Section 5.2. introducing a further subscript for the mode m. Let:

$\Delta_{od,m}$ be the link-path incidence matrix for the pair od and mode m;
Δ be the overall link-path incidence matrix, obtained by arranging side by side the blocks $\Delta_{od,m}$ corresponding to each pair od and each mode m;
c be the link cost vector, c_l ;
$\boldsymbol{g}_{od,m}^{ADD}$ be the additive path cost vector for the pair od and mode m;
$\boldsymbol{g}^{ADD}$ be the overall additive path cost vector, composed by the vectors $\boldsymbol{g}_{od,m}^{ADD}$ corresponding to each pair od and each mode m;
$\boldsymbol{g}_{od,m}^{NA}$ be the additive path cost vector for the pair od and the mode m;
$\boldsymbol{g}^{NA}$ be the overall non-additive path cost vector, composed by vectors $\boldsymbol{g}_{od,m}^{NA}$ corresponding to each pair od and each mode m;
$\boldsymbol{g}_{od,m}$ be the total path cost vector for of the pair od and the mode m;
$\boldsymbol{g}$ be the overall total path cost vector , composed by the vectors $\boldsymbol{g}_{od,m}$ corresponding to each pair od and each mode m;
$\boldsymbol{h}_{od,m}$ be the path flow vector for of the pair od and the mode m;
$\boldsymbol{h}$ be the overall path flow vector, composed by the path flows vectors $\boldsymbol{h}_{od,m}$ corresponding to each pair od and each mode m.

Generally link cost functions may be non-separable with respect to modes and other links. However, a link may be used by several modes[29]. Let

$\boldsymbol{f}^{od,m}$ be the link flow vector, with entries given by the flow on link l, $f_l^{od,m}$, corresponding to the pair od and mode m;
$\boldsymbol{f}$ be the overall link flow vector;
c be the link cost vector.

In analogy with the results presented in Section 5.2, assuming that link flows for each pair *od* and each mode *m* are measured in homogeneous units, the following holds:

$$f_l = \Sigma_m \Sigma_{od} f_l^{od,m} \tag{5.6.10}$$

The following relationships (analogous to 5.2.1-3) relate the variables introduced:

$$\boldsymbol{g}_{od,m} = \boldsymbol{g}_{od,m}^{ADD} + \boldsymbol{g}_{od,m}^{NA} = \Delta_{od,m}^{T} \boldsymbol{c} + \boldsymbol{g}_{od,m}^{NA} \qquad \forall od,m \tag{5.6.11}$$

$$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f}) \tag{5.6.12}$$

$$\boldsymbol{f} = \Sigma_m \Sigma_{od} \boldsymbol{f}^{od,m} = \Sigma_m \Sigma_{od} \Delta_{od,m} \boldsymbol{h}_{od,m} \tag{5.6.13}$$

The multi-modal supply model is expressed by the following relationship (analogous to (5.2.4):

$$\boldsymbol{g}_{od,m} = \Delta_{od,m}^{T} \boldsymbol{c}(\Sigma_m \Sigma_{od} \Delta_{od,m} \boldsymbol{h}_{od,m}) + \boldsymbol{g}_{od,m}^{NA} \qquad \forall od,m \tag{5.6.14}$$

Path choice behavior can be simulated with a random utility model, possibly different for each mode. For example, a deterministic model can be used for public transport modes, while Probit models can be specified for car and truck modes. Assuming for simplicity completely pre-trip choice behavior, let

$V_{od,m}$ be the vector of systematic utilities for paths related to the pair *od* and the mode *m*;

$p[k/odm]$ be the probability of using path *k* for a trip from the origin *o* to the destination *d* with the mode *m* (with purpose and time band not explicitly indicated),

$\boldsymbol{p}_{od,m}$ be the vector of path choice probabilities for the pair *od* and the mode *m*;

$d_{od,m}$ be the demand flow of the users between the pair *od* with mode *m*, element of the O-D matrix for mode *m*.

The following relationships (analogous to 5.2.5-6) hold between the variables introduced:

$$\boldsymbol{V}_{od,m} = -\boldsymbol{g}_{od,m} + \boldsymbol{V}^{\circ}_{od,m} \quad \forall od,m \tag{5.6.15}$$

$$\boldsymbol{h}_{od,m} = d_{od,m} \boldsymbol{p}_{od,m}(\boldsymbol{V}_{od,m}) \quad \forall od,m \tag{5.6.16}$$

where:

$V^{\circ}_{od,m}$ is a vector with elements consisting of the systematic utility components depending on any other attributes differing from path costs (such as socio-economic attributes of the users). It will be omitted in the following for simplicity of notation.

The demand flow $d_{od,m}$ for the pair od on mode m is generally defined by a system of demand models which including a mode choice model, and is therefore a function of the path choice EMPU for the various modes (analogous to 5.6.1):

$$d_{od,m} = d_{od,m}(\boldsymbol{s}) \quad \forall od,m \tag{5.6.17}$$

where:

$\boldsymbol{s}$ is the vector of the path choice EMPU, with a component $s_{od,m}$ for each pair od and each mode m.

Finally, the EMPU depends on the vector of systematic utilities (analogous to 5.2.8:

$$\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{V}) \tag{5.6.18}$$

Thus, the whole multi-mode demand model is expressed by the equation (analogous to 5.6.2):

$$\boldsymbol{h}_{od,m} = d_{od,m}(\boldsymbol{s}(-\boldsymbol{g}))\, \boldsymbol{p}_{od,m}(-\boldsymbol{g}_{od,m}) \quad \forall od,m \tag{5.6.19}$$

Combining supply and demand models, it is possible to formulate models for multi-mode equilibrium assignment analogous to the elastic demand single-mode user equilibrium assignment described in the previous sub-section. The fixed-point models are more flexible and easy to formulate, while retaining the properties described, if the mode choice model within the demand model is specified as a random utility model:

$$\boldsymbol{f}^* = \Sigma_{od,m}\, d_{od,m}(\boldsymbol{s}(-(\boldsymbol{\Delta}^T \boldsymbol{c}(\boldsymbol{f}^*) + \boldsymbol{g}^{NA}))) \, \boldsymbol{\Delta}_{od,m}\, \boldsymbol{p}_{od,m}(-(\boldsymbol{\Delta}_{od,m}{}^T \boldsymbol{c}(\boldsymbol{f}^*) + \boldsymbol{g}_{od,m}{}^{NA}))$$

The analysis of existence and uniqueness of the solutions is a simple extension of that developed in Section 5.6.1 for single-mode user equilibrium. In particular, for existence the mode choice model needs to be specified by continuous functions, while for uniqueness it needs to be specified by monotone functions, in the sense defined in Section 5.3.1. These conditions hold for additive probabilistic models expressed by continuous functions with continuous first partial derivatives.

5.7. Multi-class assignment models*

The assignment models described in the previous sections were developed assuming homogeneous users with respect to relevant behavioral models and parameters. In the following, these models are extended to deal with the case of *multi-class assignment*, i.e. assuming that users are grouped in *classes*. Users of the same class share all the behavioral characteristics such as specification, parameters and attributes of the relevant demand models, including path choice. All these features may be different than those of other classes. Users of the same class share the category and trip purpose as defined in Chapter 4[30]. User classes depend on the type of application. For example, in urban systems, classes may be identified on the basis of trip purpose, socio-economic category and the activity duration (influencing parking duration) because different travel costs (parking tolls) and different time values may be associated with these characteristics. In extra-urban systems, classes may be defined by vehicle type (auto, light and heavy commercial vehicles), trip purpose, and socio-economic characteristics, since motorway tolls, time values and path choice models may be different.

In what follows, for the sake of simplicity, reference is made to rigid demand single-mode assignment with fully pre-trip path choice behavior. The results can easily be extended to models with pre-trip/en-route choice behavior and/or with elastic demand.

The notation presented in Section 5.2 is still valid, but a further subscript i indicating the user class is considered in addition to the subscript od. Some straightforward changes in notation are described below. Let

$\Delta_{od,i}$ be the link-path incidence matrix for the pair od and class i [31];
Δ be the overall link-path incidence matrix obtained by arranging side by side the blocks $\Delta_{od,i}$ corresponding to each pair od and class i;
$d_{od,i}$ be the demand flow for the pair od and class i (for a given mode and time band);
$\boldsymbol{d}$ be the demand vector, with elements consisting of the demand flows $d_{od,i}$.

It is assumed that demand flows of each user class are measured in common units, using homogenization coefficients for users with different effects on congestion (see section 2.2). For individual modes, such as car, demand flows are expressed in vehicles per unit of time, while in the case of public modes they are expressed in passengers per unit of time.

Transport supply is simulated with a network model analogous to those described in Chapter 2. However, the cost of traversing link l can be different for users of different classes. To each link l, therefore, a cost and flow for each class can be associated. Let

f_l^i be the flow of user class i on link l;
$\boldsymbol{f}^i$ be the link flow vector for class i with entries f_l^i;
$f_l = \Sigma_i f_l^i$ be the total flow on the link l, sum of the flows corresponding to the various classes and measured in units common to the demand;
$\boldsymbol{f} = \Sigma_i \boldsymbol{f}^i$ be the vector of the total link flow with entries f_l^i
c_l^i be the cost on link l for class i;
$\boldsymbol{c}^i$ be the link cost vector for class i, with entries c_l^i.

The average cost of a path for users of class i for can be expressed as the sum of two terms: *additive path costs* with respect to class i link costs (possibly dependent on congestion), and *non-additive path costs*, which include all the specific path and/or class costs and are assumed to be independent of congestion. Let:

$\boldsymbol{g}_{od,i}^{ADD}$ be the additive path cost vector for the pair od and class i;
$\boldsymbol{g}_{od,i}^{NA}$ be the non-additive path cost vector for the pair od and class i;
$\boldsymbol{g}_{od,i}$ be the total path cost vector for the pair od and class i.

Consistency between link and path costs for each pair od and each class i, as in Chapter 2, is expressed by the following relation (analogous to 5.2.1):

$$\boldsymbol{g}_{od,i}^{ADD} = \boldsymbol{\Delta}_{od,i}^T \boldsymbol{c}^i \quad \forall od \quad \forall i$$
$$\boldsymbol{g}_{od,i} = \boldsymbol{g}_{od,i}^{ADD} + \boldsymbol{g}_{od,i}^{NA} = \boldsymbol{\Delta}_{od,i}^T \boldsymbol{c}^i + \boldsymbol{g}_{od,i}^{NA} \quad \forall od \quad \forall i \qquad (5.7.1)$$

Congestion phenomena are simulated by assuming that the cost c_l^i is a function of the class flows on the same link l, and possibly, on other links. Thus, a non-separable cost functions with respect to the classes, as well as with respect to the links, are considered. This effect is usually simulated by adopting cost functions analogous to those described in Chapter 2, in which the link congested performance attributes for each class depend on the total link flows[32]:

$$\boldsymbol{c}^i = \boldsymbol{c}^i(\boldsymbol{f}^1, \ldots, \boldsymbol{f}^i, \ldots) = \boldsymbol{c}^i(\boldsymbol{f} = \Sigma_i \boldsymbol{f}^i) \quad \forall i \qquad (5.7.2)$$

For example, the road link travel time for the car users can depend on the total flow (appropriately homogenized) of the other vehicle types (motorcycles, trucs, etc.). The cost functions of different classes, for example cars and trucks, may be different but it is assumed that they both depend on the overall link flow.

Consistency between link and path flows is expressed by the following relation (analogous to 5.2.3):

$$\boldsymbol{f}^i = \Sigma_{od} \boldsymbol{\Delta}_{od,i} \boldsymbol{h}_{od,i} \quad \forall i \qquad (5.7.3)$$

The multi-class supply model is thus described by the following equation (analogous to 5.2.4) obtained by combining equations (5.7.1-3)[33]:

$$\boldsymbol{g}_{od,i} = \boldsymbol{\Delta}_{od,i}^{T}\, \boldsymbol{c}^{i}(\Sigma_i \Sigma_{od} \boldsymbol{\Delta}_{od,i}\, \boldsymbol{h}_{od,i}) + \boldsymbol{g}_{od,i}^{NA} \quad \forall od \quad \forall i \tag{5.7.4}$$

Path choice behavior for each class i can be simulated through a random utility model, with a systematic utility equal to the opposite of the systematic path cost:

$$\boldsymbol{V}_{od,i} = -\boldsymbol{g}_{od,i,} + \boldsymbol{V}^{\circ}_{od,i} \quad \forall od \quad \forall i \tag{5.7.5}$$

where:

$\boldsymbol{V}_{od,i}$ is a vector with elements consisting of the systematic utility $V_{od,i,k}$ of path k connecting of the pair od for the class i;

$\boldsymbol{V}^{\circ}_{od,i}$ is a vector of systematic utility attributes different from those included in path costs, for simplicity of notation taken as understood in the following.

Path choice probabilities depend on the systematic utilities of alternative paths through the path choice model. Let

$\boldsymbol{p}_{od,i} = \boldsymbol{p}_{od,i}(\boldsymbol{V}_{od,i})$ be the path choice probabilities vector for pair od and class i;

$\boldsymbol{h}_{od,i,}$ be the path flow vector for the pair od and class i.

The path choice model is expressed (analogously to 5.2.6) by:

$$\boldsymbol{h}_{od,i} = d_{od,i}\, \boldsymbol{p}_{od,i}(\boldsymbol{V}_{od,i}) \quad \forall od \quad \forall i \tag{5.7.6}$$

The whole demand model is obtained by combining equations (5.2.5-6) :

$$\boldsymbol{h}_{od,i} = d_{od,i}\, \boldsymbol{p}_{od,i}(-\boldsymbol{g}_{od,i}) \quad \forall od \quad \forall i \tag{5.7.7}$$

If choice behavior on other dimensions, such as mode and destination, depends on path costs, elastic demand multi-user assignment models such as those discussed in section 5.6 are obtained. Extensions of the models to mixed pre-trip/en-route path choice behavior is analogous to those presented in section 5.5.

Multi-class assignment models can be specified by combining the supply model (5.7.4) with the demand model (5.7.7). In the following sections, multi-class assignment models will be analyzed separately for the case of congestion functions differing for different classes (*differentiated congestion*), and for the special case where congestion functions of each class are a linear transformation of a common congestion function (*undifferentiated congestion*).

5.7.1. Differentiated congestion multi-class assignment models

Differentiated congestion multi-class assignment models can be formulated with respect to path or link flows of each class. These must be consistent with the costs for each class. In the case of congested network assignment, cost functions are generally different for each class, and depend on the aggregate flow of all classes (5.7.2). The single-class assignment models described in previous paragraphs can easily be extended by considering link flows and costs per class and defining the sets of feasible path, S_h^i, and link, S_f^i, flow vectors for each class i.

Differentiated congestion multi-class uncongested network assignment models can be expressed in terms of class link flows by combining equations (5.7.1,3) with the demand model (5.7.7):

$$\boldsymbol{f}^i_{UN}(\boldsymbol{c}^i; \boldsymbol{d}_i) = \Sigma_{od}\, d_{od,i}\, \boldsymbol{\Delta}_{od,i}\, \boldsymbol{p}_{od,i}(- (\boldsymbol{\Delta}_{od,i}{}^T \boldsymbol{c}^i + \boldsymbol{g}_{od,i}{}^{NA})) \quad \forall \boldsymbol{c}_i\, \forall i \quad (5.7.8)$$

The Stochastic Uncongested Network assignment function retains the properties of continuity and monotonicity discussed in section 5.4.2, which are useful to prove existence and uniqueness of equilibrium flows as discussed below. In the case of Deterministic Uncongested Network assignment, systems of inequalities analogous to those presented in section 5.3.2 can be specified.

Differentiated congestion multi-class equilibrium assignment models are defined by combining the supply model (5.7.4) and the demand model (5.7.7). An equivalent formulation in terms of link variables can be expressed by combining the UN assignment map (5.7.8) with the cost functions (5.7.2). Extension to elastic or multi-modal demand assignment (section 5.6) or to mixed pre-trip/en-route path choice behavior (section 5.5) is relatively straightforward.

Stochastic multi-class equilibrium can be formulated with fixed-point models analogous to those described in the previous sections, while deterministic multi-class user equilibrium can also be formulated with variational inequality models.

Existence conditions of the multi-class equilibrium configurations require the continuity of the cost functions, $\boldsymbol{c}^i()$, for each class i, with respect to the flows of the *various classes*, $\boldsymbol{f}^1,\ldots,\boldsymbol{f}^i, \ldots,$. Note that continuity with respect to the total flows, $\boldsymbol{f}$, also ensures the continuity with respect to the class flows, $\boldsymbol{f}^i$, and therefore the existence of the equilibrium configurations.

Uniqueness conditions of multi-class equilibrium configurations require the monotonicity of the cost functions, $\boldsymbol{c}^i = \boldsymbol{c}^i()$, for each class i with respect to the flows of the various classes, $\boldsymbol{f}^1,\ldots,\boldsymbol{f}^i, \ldots,$ defined by the following condition:

$$\Sigma_i\, [\boldsymbol{c}^i(\boldsymbol{f}^1,\ldots,\boldsymbol{f}^i, \ldots) - \boldsymbol{c}^i(\boldsymbol{y}^1,\ldots,\boldsymbol{y}^i, \ldots)]^T\, (\boldsymbol{f}^i - \boldsymbol{y}^i) > 0$$
$$\forall (\boldsymbol{f}^1,\ldots,\boldsymbol{f}^i, \ldots) \neq (\boldsymbol{y}^1,\ldots,\boldsymbol{y}^i, \ldots) : \boldsymbol{f}^i, \boldsymbol{y}^i \in S_f^i \;\; \forall i$$

or

$$\Sigma_i\, [\boldsymbol{c}^i(\Sigma_j \boldsymbol{f}^j) - \boldsymbol{c}^i(\Sigma_j \boldsymbol{y}^j)]^T\, (\boldsymbol{f}^i - \boldsymbol{y}^i) > 0 \quad (5.7.9)$$
$$\forall (\boldsymbol{f}^1,\ldots,\boldsymbol{f}^i, \ldots) \neq (\boldsymbol{y}^1,\ldots,\boldsymbol{y}^i, \ldots) : \boldsymbol{f}^i, \boldsymbol{y}^i \in S_f^i \;\; \forall i$$

It should be noted that in general the strict monotonicity of the class cost functions with respect to class flows, defined by (5.7.9), and therefore the uniqueness of the equilibrium configuration, is not ensured by the strict monotonicity of the class cost functions with respect to the total link flows, as defined by the following different condition:

$$[\boldsymbol{c}^i(\boldsymbol{f}) - \boldsymbol{c}^i(\boldsymbol{x})]^{\mathrm{T}} (\boldsymbol{f} - \boldsymbol{x}) > 0 \;\forall i$$
$$\forall \boldsymbol{f} = \Sigma_j \boldsymbol{f}^j \neq \boldsymbol{x} = \Sigma_j \boldsymbol{x}^j : \boldsymbol{f}^j, \boldsymbol{x}^j \in S_f^i \quad \forall i$$

or

$$[\boldsymbol{c}^i(\Sigma_j \boldsymbol{f}^j) - \boldsymbol{c}^i(\Sigma_j \boldsymbol{x}^j)]^{\mathrm{T}} \Sigma_j (\boldsymbol{f}^j - \boldsymbol{x}^j) > 0 \;\forall i \qquad (5.7.10)$$
$$\forall \Sigma_j \boldsymbol{f}^j \neq \Sigma_j \boldsymbol{x}^j : \boldsymbol{f}^j, \boldsymbol{x}^j \in S_f^i \quad \forall i$$

In fact, the sum over the index i of inequalities (5.7.10) does not necessarily imply the condition (5.7.9)[34].

It should also be noted that the symmetry of the cost function Jacobian necessary for the formulation of the deterministic equilibrium with optimization models, relates not only to the effect of flow on a link on costs of different links but also of flow of a class on the cost of other classes for the same link. Similarly, separability of cost functions requires that the cost of class i on link l, c_l^i, depends only on the flow on the same link l, f_l^i, of the same class. This second condition is almost never satisfied in applications.

In general, the problem of differentiated congestion multi-class equilibrium assignment can be formulated by extending the corresponding single-class models. However the (sufficient) uniqueness conditions are seldom satisfied.

5.7.2. Undifferentiated congestion multi-class assignment models

In *undifferentiated congestion multi-class assignment* it is assumed that the class cost functions can be expressed as a linear transformation of a cost function common to all the classes and dependent on total link flows. These costs are called *reference costs*. Therefore, multi-class equilibrium assignment can be formulated in terms of total flows and reference link costs. Under the assumptions made, expression (5.7.2) for the link cost function becomes:

$$c_l^i = c^i{}_l(\boldsymbol{f}) = \gamma_i \, \bar{c}_l(\boldsymbol{f}) + c^i{}_{0,l} \quad \forall i \qquad (5.7.11)$$

where:

$\bar{c}_l = \bar{c}_l(\boldsymbol{f})$ is the reference cost function of link l;

$\gamma_i \geq 0$ is the ratio (assumed independent of the link) between the link cost for class i and the reference cost, if $\gamma_i = 0$ the class i costs are uncongested;

$c^i{}_{0,l}$ is the cost of link l specific to class i, assumed to be independent of congestion.

All costs are assumed to be expressed in units homogeneous with the utility (through relevant coefficients). The reference cost function, $\bar{c}_l(\boldsymbol{f})$ can represent disutility related to the average travel time, while c_{ol} the disutility connected to monetary costs, possibly different for different classes and/or with different substitution coefficients. The coefficients γ_i can express the ratios between class-specific and average travel times.

Using expression (5.7.11), the consistency between link and path costs is expressed for each pair *od* and class *i* by the following relation:

$$\boldsymbol{g}_{od,i} = \boldsymbol{\Delta}_{od,i}^{T} (\gamma_i \, \bar{\boldsymbol{c}} + \boldsymbol{c}^{i}_{0}) + \boldsymbol{g}_{od,i}^{NA} \quad \forall i \quad \forall od$$
$$\boldsymbol{g}_{od,i} = \gamma_i \, \boldsymbol{\Delta}_{od,i}^{T} \, \bar{\boldsymbol{c}} + \boldsymbol{\Delta}_{od,i}^{T} \boldsymbol{c}^{i}_{0} + \boldsymbol{g}_{od,i}^{NA} \quad \forall i \quad \forall od$$

where:

$\bar{\boldsymbol{c}}$ is the vector of reference link costs ;
$\boldsymbol{c}^{i}_{0}$ is the vector of class *i* specific link costs;
$\boldsymbol{g}_{od,i}^{NA}$ is the vector of non-additive path costs for the pair *od* and class *i*;
$\boldsymbol{g}_{od,i}$ is the total path cost vector for the pair *od* and class *i*, in utility units.

The average cost of a path between the pair *od* for a user of class *i* therefore consists of two components:

- *additive (and generic) costs*, the sum of reference link costs, possibly dependent on congestion, given by $\gamma_i \, \boldsymbol{\Delta}_{od,i}^{T} \bar{\boldsymbol{c}}$;
- congestion-independent path costs consisting of:
  - *(additive and) class specific costs*, sum of class-specific link costs, given by $\boldsymbol{\Delta}_{od,i}^{T} \boldsymbol{c}_{i,0}$;
  - *non-additive costs*, which cannot be expressed as the sum of link costs, however defined, given by $\boldsymbol{g}_{od,i}^{NA}$.

Let

$\boldsymbol{g}_{od,i}^{SPNA} = \gamma_i \boldsymbol{\Delta}_{od,i}^{T} \boldsymbol{c}^{i}_{0} + \boldsymbol{g}_{od,i}^{NA}$ be the vector of specific and/or non-additive path costs for the pair *od* and class *i*

A relationship between link and path costs analogous to (5.7.1) can be formulated:

$$\boldsymbol{g}_{od,i} = \gamma_i \, \boldsymbol{\Delta}_{od,i}^{T} \, \bar{\boldsymbol{c}} + \boldsymbol{g}_{od,i}^{SPNA} \quad \forall_{od} \quad \forall i \qquad (5.7.12)$$

The undifferentiated congestion multi-class supply model is thus described by the following relation obtained by combining eqn. (5.7.3) with equations (5.7.11,12) and the reference cost functions given by (5.2.2):

$$\boldsymbol{g}_{od,i} = \gamma_i \, \boldsymbol{\Delta}_{od,i}^{T} \, \bar{\boldsymbol{c}} \, (\Sigma_i \, \Sigma_{od} \boldsymbol{\Delta}_{od,i} \boldsymbol{h}_{od,i}) + \boldsymbol{g}_{od,i}^{SPNA} \quad \forall od \quad \forall i \qquad (5.7.13)$$

Path choice behavior is simulated by a random utility model, expressed by (5.7.6), in which the systematic utility of a path is equal to the opposite of the path average cost for class i, as expressed in the relation (5.7.5). In the case of a Logit path choice model, the parameter γ_i cannot be identified separately from the parameter θ. In the case of a deterministic path choice model, γ_i is not relevant, since it does not change the maximum systematic utility alternative, i.e. the minimum cost path[35].

Under the given assumptions, undifferentiated congestion multi-class assignment models can therefore be defined with respect to total path or link flows, consistent with reference link costs and the interaction between classes. The considerations made in the previous sections are still valid. In particular, the sets of the feasible path S_F and link S_f flows are defined as in Section 5.2.

Undifferented congestion uncongested network multi-class assignment models are expressed by:

$$\boldsymbol{f}_{UN}(\bar{\boldsymbol{c}}\,;\boldsymbol{d},\gamma) = \Sigma_{od,i}\, d_{od,i}\, \Delta_{od,i}\, \boldsymbol{p}_{od,i}(-\gamma_i \Delta_{od,i}^{T}\, \bar{\boldsymbol{c}} - \boldsymbol{g}_{od,i}^{SPNA}) \qquad (5.7.14)$$

The stochastic uncongested network assignment function retains the properties of continuity and monotonicity discussed in Section 5.3.2 if the coefficients γ_i are non-negative. In the case of deterministic assignment, systems of inequalities analogous to those presented in Section 5.3.2 can be developed.

Undifferentiated congestion equilibrium multi-class assignment models are defined by the system of equations obtained by combining the supply model (5.7.13) and the demand model (5.7.7). An equivalent formulation in terms of total link flows, $\boldsymbol{f}$, and of reference link costs, $\bar{\boldsymbol{c}}$, can be expressed by the system of equations obtained by combining the UN assignment map (5.7.14) with the reference cost functions given by (5.2.2). Stochastic or deterministic user equilibrium assignment can be formulated with fixed-point or variational inequality models respectively, analogous to the models presented in the previous paragraphs. Continuity and monotonicity of the link reference cost functions are required for the existence and the uniqueness of the equilibrium solution:

$$[\bar{\boldsymbol{c}}(\boldsymbol{f}') - \bar{\boldsymbol{c}}(\boldsymbol{f}'')]^{T}(\boldsymbol{f}' - \boldsymbol{f}'') > 0 \quad \forall \boldsymbol{f}' \neq \boldsymbol{f}'' \in S_f$$

It can easily be deduced that the parameters γ_i, assumed non-negative, do not alter the existence and uniqueness conditions of equilibrium solutions. However, they influence the value of SUE solution, while, as noted earlier, they have no influence in the case of DUE assignment.

Finally, it must be noted that in stochastic equilibrium, once the equilibrium total link flows $\boldsymbol{f}^*$ are known, it is possible to compute equilibrium reference costs $\bar{\boldsymbol{c}}^*$ and therefore class-specific link, $\boldsymbol{c}^i$, and path, $\boldsymbol{g}_i$, costs. From these costs, class-specific path flows $\boldsymbol{h}_i$ and consequently link flows $\boldsymbol{f}^i$ can be obtained:

$$f^i = \Sigma_{od} \Delta_{od,i} p_{od,i}(-\gamma_i \Delta_{od,i}^T \bar{c}(f) - g_{od,i}^{SPNA}) \quad \forall i$$

Existence and uniqueness of stochastic equilibrium total link flows ensure the existence and uniqueness of class specific flows. On the other hand, in the case of deterministic models, several class specific link flows could be associated with the same link cost vector if there were several minimum cost paths. Thus, in the case of deterministic multi-class equilibrium, the existence of total equilibrium link flows ensures the existence of class flows, but the uniqueness of total link flows does not guarantee the uniqueness of class-specific link flows. In this case, to guarantee the uniqueness of class link flows, an explicit formulation in terms of class flows is necessary as in the case of differentiated congestion assignment.

5.8. Inter-period Dynamic Process assignment models*

User equilibrium models define a priori the relevant state of the system as the one in which average demand and costs are mutually consistent.

Alternatively, dynamic process assignment models simulate the evolution of the system over a sequence of similar periods, days or their parts[36], and the possible convergence of the system to a stable condition. For this reason, dynamic process models are also known as *"non-equilibrium" models*. As was noted in Chapter 1, this type of dynamic is known as *inter-period* or *day-to-day dynamics*. Dynamic process models are based on (non-linear) dynamic systems theory or on stochastic processes theory, according to whether the state of the system is described by deterministic or random variables.

Dynamic process models, which are a sector of growing research interest, can be seen as a generalization of equilibrium models since they simulate the convergence of the supply-demand system towards possibly different equilibrium states and the transient states visited due to modifications of supply and/or demand. Furthermore, under some rather mild assumptions, equilibrium configurations of the system described in previous sections can be modeled as attractors of the system, i.e. states in which the evolution of the system stops. Finally, the dynamic approach allows analysis of the stability of equilibrium configurations and provides a complete statistical description of the system's evolution.

In general, the specification of a dynamic process model requires a more detailed simulation of users' behavior and in particular the explicit modeling of two phenomena (Fig. 5.8.1), which are not relevant in the equilibrium approach:

- the users' choice updating behavior, i.e. how present choices are influenced by the choice made on previous days, including phenomena such as habit (*choice updating model*);
- the users' learning and forecasting mechanisms, i.e. how experience and the transport cost information history influence present choices, including phenomena such as memory and information spreading (*utility updating model*).

5.8.1. Definitions, assumptions and basic equations

This section presents the basic relationships defining a dynamic process assignment model. For the sake of clarity, rigid demand single-mode single-class[37] assignment is considered. It is also assumed that path choice behavior is probabilistic and fully pre-trip. Some of the variables presented in Section 5.2 should be redefined in order to associate them with the evolution of the system over a sequence of reference periods (inter-period or day-to-day dynamics). Let

t be the generic reference period, assumed for the sake of simplicity as the day;
Δ_{od} be the link path incidence matrix for the pair od, assumed to be independent of the day;
Δ be the total link path incidence matrix;
$\boldsymbol{h}_{od}^{t}$ be the od vector of the path flows on day t;
$\boldsymbol{h}^{t}$ be the total vector of the path flows on day t;
$\boldsymbol{f}^{t}$ be the vector of the link flows on day t;
$\boldsymbol{r}_{n}^{t}$ be the vector of n-th link performance attributes on day t;
$\boldsymbol{c}^{t}$ be the vector of (average) link costs on day t;
$\boldsymbol{g}_{od}^{t}$ be the od vector of (average) path costs on day t;
$\boldsymbol{g}$ be the total vector of the (average) path costs on day t.

5.8.1.1. Supply model

Supply is simulated by applying the relations (5.2.1-3) to costs and flows on day t. Ignoring non-additive path costs for simplicity, $\boldsymbol{g}_{od}^{NA} = \boldsymbol{0}$, it follows that:

$$g_{od}{}^{t} = \Delta_{od}{}^{T}\, \boldsymbol{c}^{t} \qquad (5.8.1a)$$

$$\boldsymbol{g}^{t} = \Delta^{T}\, \boldsymbol{c}^{t} \qquad (5.8.1b)$$

$$\boldsymbol{c}^{t} = \boldsymbol{c}(\boldsymbol{f}^{t}) \qquad (5.8.2)$$

$$\boldsymbol{f}^{t} = \Sigma_{od}\, \Delta_{od}\, \boldsymbol{h}_{od}^{t} \qquad (5.8.3a)$$

$$\boldsymbol{f}^{t} = \Delta\, \boldsymbol{h}^{t} \qquad (5.8.3b)$$

Combining equations (5.8.1-3), the following relation between path costs, $\boldsymbol{g}^{t}$, and the path flows, $\boldsymbol{h}^{t}$, on day t is obtained:

$$\boldsymbol{g}_{od}{}^{t} = \Delta_{od}{}^{T}\, \boldsymbol{c}(\Sigma_{od}\, \Delta_{od}\, \boldsymbol{h}_{od}^{t}) \quad \forall od \qquad (5.8.4a)$$

$$\boldsymbol{g}^{t} = \Delta^{T}\, \boldsymbol{c}(\Delta\, \boldsymbol{h}^{t}) \qquad (5.8.4b)$$

Equations (5.8.4) define the *supply model* corresponding to day t. It is readily apparent that the relation (5.8.4) is analogous to (5.2.4) defining the supply model in the static case.

5.8.1.2. Demand model

The simulation of day-to-day dynamic path choice behavior requires extending the static demand model relations (5.2.5-7). In particular, the relationships between the costs on different days and the attributes influencing the users' choices, as well as the choice updating mechanisms on subsequent days, must be made explicit. Let

$d_{od} \geq 0$ be the demand flow for the users of the pair *od*, assumed to be independent of the day for the sake of simplicity (consistent with the rigid demand hypothesis);

$\boldsymbol{d}$ be the demand vector, whose components are the demand values d_{od} for the each O-D pair;

V_{od}^{t} be the vector systematic path utilities forecasted on day t by the users of the pair *od*;

$\boldsymbol{V}^{t}$ be the total vector of systematic path utilities forecast on day t.

The *utility updating model* simulates the way in which perceived utilities on day t are influenced by utilities and costs on previous days (and possibly by others sources of information). In principle, for disaggregate assignment models, the updating of individual user utilities can be modelled expressing the dependence of perceived utilities for all paths k on day t, $U_k^{i,t}$, on previous the perceived utilities on previous days and actual costs. This can be expressed symbolically as

$$U_{od}^{i,.t} = U(U_{od}^{i,t-1}\ U_{od}^{i,t-2}.., g_{od}^{t-1}, g_{od}^{t-2}.....)$$

This model, however, is not applicable to aggregate assignment. Furthermore, it would be complex to specify choice models based on random utility theory given the serial correlation of the day t random residuals on those of previous days. The models proposed in the literature are special cases; they assume that utility updating is carried out on average (systematic) utilities through a function known as a *filter*, $\boldsymbol{V}()$. The filter is a generalization of systematic utility functions defined in the static case by the relation (5.2.5):

$$V_{od}^{t} = V_{od}(V_{od}^{t-1}, \boldsymbol{g}_{od}^{t-1}, V_{od}^{t-2}, \boldsymbol{g}_{od}^{t-2}, \ldots) \quad \forall od$$

$$\boldsymbol{V}^{t} = \boldsymbol{V}(\boldsymbol{V}^{t-1}, \boldsymbol{g}^{t-1}, \boldsymbol{V}^{t-2}, \boldsymbol{g}^{t-2}, \ldots)$$

For the sake of simplicity of in the following, it is assumed that expected (or predicted) average utilities on day t depend only on the actual costs, $\boldsymbol{g}^{t-1}$, and on the expected utilities, $\boldsymbol{V}^{t-1}$, on the previous day, t-1:

$$V_{od}^{t} = V_{od}(V_{od}^{t-1}, \boldsymbol{g}_{od}^{t-1}) \quad \forall od \qquad (5.8.5a)$$

$$\boldsymbol{V}^{t} = \boldsymbol{V}(\boldsymbol{V}^{t-1}, \boldsymbol{g}^{t-1}) \qquad (5.8.5b)$$

Note that under this assumption, the actual costs, g_{od}^{t-2}, on the days previous to t-1, still influence the choice behavior on day t since they influence the expected utility V^{t-1} on the previous day t-1.

A simple example of a utility updating model is defined by an *exponential filter* in which the expected utility on day t is expressed by a convex combination of the expected utility on day t-1, V^{t-1}, and the opposite of the actual path costs on day t-1, $-g^{t-1}$, as defined by the supply model (5.8.4). Relation (5.8.5) becomes:

$$V^t = -\beta g^{t-1} + (1-\beta)V^{t-1} \qquad \forall od \tag{5.8.5c}$$

where:

$\beta \in]0,1]$ is the average weight attributed by the users to the actual costs on day t-1; if $\beta = 1$ the expected utility is equal to the negative actual cost on day t-1, and the costs on previous days do not influence users' behavior. This parameter is usually assumed to be independent of the day and may be different for different classes of users.

Given the linear relationship between link and (additive) path costs, the exponential filter can be applied also to link costs:

$$x^t = \beta c^{t-1} + (1-\beta)x^{t-1}$$

where x^t is the vector of expected link costs on day t. In this case, the expected path utilities on day t are given by[38]:

$$V_{od}^t = -\Delta_{od}^T x^t$$
$$V^t = -\Delta^T x^t$$

The *choices updating model* simulates the way in which day t choices are influenced by the choices made on previous days. The most general approach can be expressed by a square matrix R^t, known as a *conditional choice matrix*, which has a number of rows and columns equal to the number of paths. The elements $r_{k,j}^t \in [0,1]$ are the conditional path choice fractions, i.e. the fraction of users choosing the path k on day t given the path j chosen on day t-1. Since $r_{k,j} = 0$ if the paths k and j do not connect the same pair od, the following holds: $\Sigma_{k \in Iod} r_{k,j}^t = 1 \quad \forall j \in I_{od}$.

The path flow vector on day t, h^t, can therefore be expressed as the product of the conditional choice matrix, R^t, and the path flow vector on the previous day t-1, h^{t-1}:

$$h_k^t = \Sigma_{j \in Iod} r_{k,j}^t h_j^{t-1} \qquad \forall k \in I_{od} \qquad \forall od$$
$$h_{od}^t = R_{od}^t h_{od}^{t-1} \qquad \forall od$$
$$h^t = R^t h^{t-1}$$

Note that the path flow vector on day t is feasible, $\boldsymbol{h}^t \in S_h$, if the path flow vector on the previous day is feasible, $\boldsymbol{h}^{t-1} \in S_h$, (i.e. if it is non-negative and satisfies the demand conservation constraint).

The elements of the conditional choice matrix (or rather their average values), $\boldsymbol{R}^t$, can be simulated with a random utility model as a function of the expected utilities on day t (and possibly of other days and/or of other attributes not expressed here). In this way, a generalization of the path choice models adopted in the static case is obtained:

$$\boldsymbol{R}_{od}^{\,t} = \boldsymbol{R}_{od}(\boldsymbol{V}_{od}^{\,t})$$
$$\boldsymbol{R}^t = \boldsymbol{R}(\boldsymbol{V}^t)$$

Combining the two previous relationships, a generalization of the static-case relation (5.2.6) is obtained:

$$\boldsymbol{h}_{od}^{\,t} = \boldsymbol{R}_{od}(\boldsymbol{V}_{od}^{\,t})\, \boldsymbol{h}_{od}^{\,t-1} \quad \forall od \qquad (5.8.6a)$$
$$\boldsymbol{h}^t = \boldsymbol{R}(\boldsymbol{V}^t)\boldsymbol{h}^{t-1} \qquad (5.8.6b)$$

A simple example of a choice updating model for the simulation of the conditional choice matrix is the *exponential filter* model. This model assumes that each day some users repeat the choices made the previous day, while the others reconsider (though do not necessarily change) their choices with a probability independent of the choice made on the previous day:

$$r_{kk}^{\,t} = \alpha\, p_k^{\,t} + (1-\alpha) \qquad \forall k \in I_{od} \quad \forall od$$
$$r_{kj}^{\,t} = \alpha\, p_j^{\,t} \qquad \forall j \neq k, j \in I_{od} \quad \forall k \in I_{od} \quad \forall od$$

where:

$p_k^{\,t} \in]0,1]$ is the probability that a user reconsidering the choice made the previous day, t-1, chooses the path $k \in I_{od}$ on day t;

$\alpha \in]0,1]$ is the probability that a user reconsiders the choice made the previous day, therefore $(1-\alpha)$ is the probability that the previous day's choice is repeated; if $\alpha = 1$ all the users reconsider their previous day choices; this parameter is usually assumed to be independent of the day[39] but may differ by user class.

Under this model, it follows that:

$$h_k^{\,t} = \Sigma_{j \in Iod}\ \alpha\, p_k^{\,t}\, h_j^{\,t-1} + (1-\alpha) h_k^{\,t-1} = \alpha\, p_k^{\,t}\, \Sigma_{j \in Iod}\, h_j^{\,t-1} + (1-\alpha) h_k^{\,t-1} \quad \forall k \in I_{od}\ \forall od$$

Since $d_{od} = \Sigma_{j \in Iod}\, h_k$, it yields:

$$\boldsymbol{h}_{od}^{\,t} = \alpha\, d_{od} \boldsymbol{p}_{od}^{\,t} + (1-\alpha)\boldsymbol{h}_{od}^{\,t-1}$$

The path choice probability, p_k^t, is usually obtained with one of the path choice models described in section 4.2.5, $\boldsymbol{p}_{od}^t = \boldsymbol{p}_{od}(\boldsymbol{V}_{od}^t)$. The relation (5.8.6) therefore becomes (cfr. 5.2.6):

$$\boldsymbol{h}_{od}^t = \alpha\, d_{od}\boldsymbol{p}_{od}(\boldsymbol{V}_{od}^t) + (1-\alpha)\boldsymbol{h}_{od}^{t-1} \qquad (5.8.6c)$$

By combining the two recursive equations (5.8.5) and (5.8.6), we get a relationship between the path flows $\boldsymbol{h}^t$ on day t and paths costs $\boldsymbol{g}^{t-1}$ on the day t-1 which defines the *demand model* corresponding to day t,:

$$\boldsymbol{h}_{od}^t = \boldsymbol{R}_{od}(\boldsymbol{V}_{od}(\boldsymbol{V}_{od}^{t-1}, \boldsymbol{g}_{od}^{t-1}))\,\boldsymbol{h}_{od}^{t-1} \quad \forall od \qquad (5.8.7a)$$

$$\boldsymbol{h}^t = \boldsymbol{R}(\boldsymbol{V}(\boldsymbol{V}^{t-1}, \boldsymbol{g}^{t-1})\,\boldsymbol{h}^{t-1} \qquad (5.8.7b)$$

This relation is a generalization of the static case (5.2.7). If exponential filters are adopted to formulate utility and choice updating models, expression (5.8.7a) becomes:

$$\boldsymbol{h}_{od}^t = \alpha\, d_{od}\boldsymbol{p}_{od}(-\beta\boldsymbol{g}_{od}^{t-1} + (1-\beta)\bar{\boldsymbol{V}}_{od}^{t-1}) + (1-\alpha)\boldsymbol{h}_{od}^{t-1} \qquad (5.8.7c)$$

5.8.1.3. Approaches to Dynamic Process modeling

The recursive equations (5.8.7) defining the choice model, combined with the recursive equations (5.8.5) for the updating of the expected utilities and the supply model (5.8.4) identify a dynamic process model. The state of the system on day t is defined by the vectors of predicted systematic utilities, $\boldsymbol{V}^t$, and by the path flows, $\boldsymbol{h}^t$, describing the combined results of the utility and choice updating models as a function of the state on the previous day[40]:

$$\boldsymbol{V}^t = \boldsymbol{V}(\boldsymbol{V}^{t-1}, \Delta^t\boldsymbol{c}(\Delta\boldsymbol{h}^{t-1})) \qquad (5.8.8)$$

$$\boldsymbol{h}^t = \boldsymbol{R}(\boldsymbol{V}^t)\,\boldsymbol{h}^{t-1} \qquad (5.8.9)$$

The set of feasible states S, known as *state space*, is defined by the vectors of expected path utilities, $\boldsymbol{V}^t \in R^n$, and the feasible path flows, $\boldsymbol{h}^t \in S_h$, or $S = S_h \times R^n$.

Given an initial state, the recursive equations (5.8.8-9) define a dynamic process model (Fig. 5.8.1). If the vectors of path flows, $\boldsymbol{h}^t$, and predicted utilities, $\boldsymbol{V}^t$, are modeled as deterministic variables, a *deterministic process* model results, while if they are modeled as random variables a *stochastic process* model is obtained (Fig. 5.8.2). A deterministic process model can also be interpreted as a process approximating the expected values of the corresponding stochastic process.

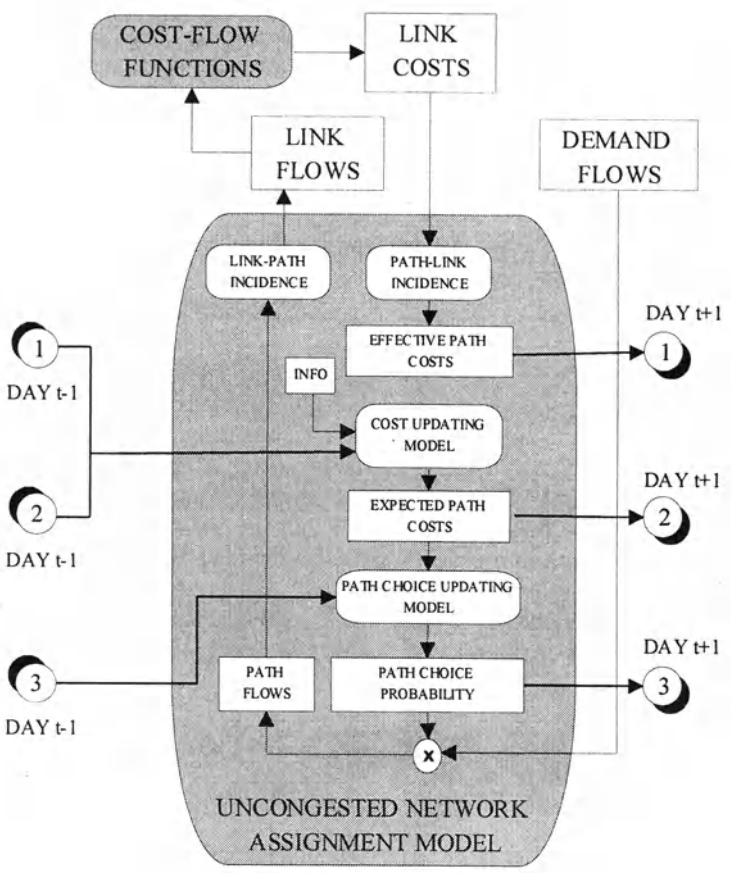


Fig. 5.8.1 – Schematic representation of Dynamic Process (DP) assignment models.

Note that the terms stochastic and deterministic have a different meaning when referring to the dynamic process versus the path choice models in assignment. In the former case, they relate to the actual representation of the system, i.e. to the assumption made by the analyst to represent the state variables. In the second case, they relate to the assumptions made in modeling path choices, i.e. the presence of a random residual in the utility functions, and therefore the form of path choice models. Equilibrium models, either deterministic or stochastic, refer to a "deterministic" representation of the system.

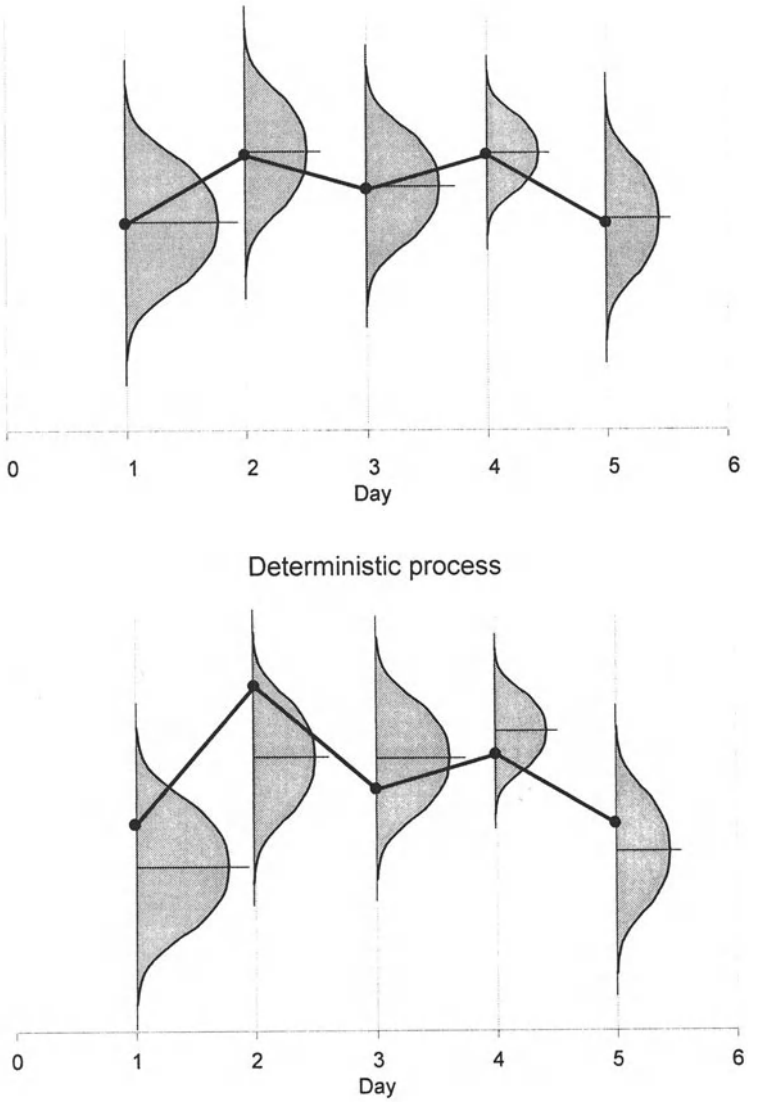


Fig. 5.8.2 – Graphic representation of deterministic and stochastic process models

5.8.2. Deterministic Process models

Deterministic process models derive from the assumption that the path flows and utilities predicted on day t are represented by deterministic variables, i.e. that the flows and utilities coincide with their average values. The system evolution over time, in terms of path flows and utilities, is defined by the recursive equations (5.8.8-9). A model of this type allows the analysis of system convergence and, if this is the case, toward which subset of the state space. Recursive equations (5.8.8-9) define the *transition function* ψ relating the state on the day t to the state on the previous day t-1:

$$(\boldsymbol{h}^t, \boldsymbol{V}^t) = \psi(\boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}) \tag{5.8.10}$$

According to the theory of (non-linear) dynamic systems, any proper subset $A \subset S$ of the state space $S \subseteq R^N$, with a dimension strictly smaller than the dimension N of S, [41] is called an *attractor* if:

- the system cannot evolve toward a state outside the attractor starting from a state inside it;
- the attractor is properly contained in another subset $B \subseteq S$ (called the *basin* of A), such that if the initial state is contained in B the final state tends to be contained in A;
- A is minimal in the sense that it does not properly contain other attractors.

In other words, if the initial state is sufficiently close to the attractor, the system evolves towards it and, once reached, does not leave. Note that a system may have several attractors, each with its own basin[42]. A classification of attractors is given in Fig. 5.8.3 (examples are given in Fig. 5.8.4).

	TYPES OF ATTRACTOR A	number of points in A	dimension of A (< N)
NON CAOTIC	FIXED POINT the system always occupies the same point	1	0
Evolutions starting from near states remain close	K-PERIODIC the system periodically occupies k points	*k*	0
	QUASI-PERIODIC the system moves on a torus (or a set of tori)	∞	integer
CAOTIC	A-PERIODIC the system moves in a fractal set	∞	non-integer

Fig. 5.8.3 – Attractors of a deterministic dynamic process

If *fixed-point* states $(\boldsymbol{h}^*,\boldsymbol{V}^*)$ (not necessarily attractors) are reached, the evolution of the system stops:

$$(\boldsymbol{h}^l, \boldsymbol{V}^l) = (\boldsymbol{h}^{l-1}, \boldsymbol{V}^{l-1}) = (\boldsymbol{h}^*,\boldsymbol{V}^*)$$

that is

$$(\boldsymbol{h}^*,\boldsymbol{V}^*) = \psi(\boldsymbol{h}^*,\boldsymbol{V}^*) \tag{5.8.11}$$

This condition, combined with (5.8.8-9), leads to:

$$\boldsymbol{V}^* = \boldsymbol{V}(\boldsymbol{V}^*, \Delta' \boldsymbol{c}(\Delta \boldsymbol{h}^*))$$
$$\boldsymbol{h}^* = \boldsymbol{R}(\boldsymbol{V}^*)\, \boldsymbol{h}^*$$

In general, fixed-point states depend on the utility and choice updating models (and are different from equilibrium states).

An example of a deterministic process is obtained by adopting exponential filter specifications for the utility and choice updating models presented in Section 5.8.1.2. In this case, equation (5.8.5c) can be reformulated for all the pairs *od*:

$$\boldsymbol{V}^l = -\beta \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^{l-1}) + (1-\beta)\boldsymbol{V}^{l-1} \tag{5.8.12}$$
$$\boldsymbol{h}^l = \alpha \boldsymbol{P}(\boldsymbol{V}^l)\boldsymbol{d} + (1-\alpha)\boldsymbol{h}^{l-1} \tag{5.8.13}$$

Similarly, the model can be expressed in terms of link flows and expected costs:

$$\boldsymbol{x}^l = \beta \boldsymbol{c}(\boldsymbol{f}^{l-1}) + (1-\beta)\boldsymbol{x}^{l-1} \tag{5.8.14}$$
$$\boldsymbol{f}^l = \alpha \boldsymbol{f}_{SUN}(\boldsymbol{x}^l) + (1-\alpha)\boldsymbol{f}^{l-1} \tag{5.8.15}$$

Fixed-point states of the process defined by (5.8.12-13) are given by:

$$\boldsymbol{g}^* = \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^*) \tag{5.8.16}$$
$$\boldsymbol{h}^* = \boldsymbol{P}(-\boldsymbol{g}^*)\boldsymbol{d} \tag{5.8.17}$$

and for the process defined by (5.8.14-16) in terms of link flows and costs by:

$$\boldsymbol{c}^* = \boldsymbol{c}(\boldsymbol{f}^*) \tag{5.8.18}$$
$$\boldsymbol{f}^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}^*) \tag{5.8.19}$$

In this case, it can be immediately verified that the formulations in terms of path or link variables are equivalent. Furthermore, the fixed-point states coincide with the stochastic user equilibrium states defined in section 5.4, and the conditions of existence and uniqueness discussed still hold. Note also that the definition, existence, and uniqueness of fixed-points do not depend on the parameters α and β, which specify the choice and utility updating filters respectively[43].

Examples of the evolution of the transportation system depicted in Fig. 5.4.2 for different values of the parameters are given in Fig. 5.8.4. It should be noted that for

some values of the parameters, link flows converge to a fixed-point state which coincides with the SUE configuration.

By applying the theory of non-linear dynamic systems, it is possible to define conditions ensuring that a fixed-point state is (*locally*) *stable*, i.e. it has an attraction basin which is (a subset of) the state space *S*. In particular, if the transition function of any deterministic process model $(\boldsymbol{h}^l, \boldsymbol{V}^l) = \psi(\boldsymbol{h}^{l-1}, \boldsymbol{V}^{l-1})$ is continuous and differentiable with continuous Jacobian, $\boldsymbol{Jac}[\psi(\boldsymbol{h}^{l-1}, \boldsymbol{V}^{l-1})]$, a fixed point $(\boldsymbol{h}^*, \boldsymbol{V}^*)$ is stable if all the eigenvalues(44) of the Jacobian at the fixed-point $\boldsymbol{Jac}[\psi(\boldsymbol{h}^*,\boldsymbol{V}^*)]$ have absolute values less than one. The transition function Jacobian, and therefore its eigenvalues, depend on the utility and choice updating models, which therefore influence the stability of a fixed-point state.

To facilitate the comparison with equilibrium, the following analysis considers the model formulated in terms of link flows and costs (5.8.14-15). Assume also that the transition function $(\boldsymbol{f}^l, \boldsymbol{c}^l) = \psi(\boldsymbol{f}^{l-1}, \boldsymbol{c}^{l-1})$ is continuous and differentiable with a continuous Jacobian $\boldsymbol{Jac}[\psi(\boldsymbol{f}^{l-1}, \boldsymbol{c}^{l-1})]$. In this case, the dynamic system is defined by $2n_L$ variables where n_L is the number of links and the Jacobian has $2n$ eigenvalues, two for each link l, denoted by λ_l and λ_{n+l}. Under these assumptions, a fixed-point state defined by (5.8.18-19) is stable if all the eigenvalues of the Jacobian calculated at the fixed-point $\boldsymbol{Jac}[\psi(\boldsymbol{f}^*,\boldsymbol{c}^*)]$ have absolute values less than one:

$$|\lambda_l^*| < 1 \quad \forall l$$
$$|\lambda_{n+l}^*| < 1 \quad \forall l$$

This condition constrains the eigenvalues to the interior of a unit-radius circle on the complex plane (Argand plane).

The Jacobian $\boldsymbol{Jac}[\psi(\boldsymbol{f}, \boldsymbol{c})]$ of the transition function $(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{f}, \boldsymbol{c})$ for the model (5.8.14-15) at the point $(\boldsymbol{f}, \boldsymbol{c})$ is given by:

$$\boldsymbol{Jac}[\psi(\boldsymbol{f},\boldsymbol{c})] = \begin{array}{|c|c|} \hline (1-\beta)\boldsymbol{I} & \beta\boldsymbol{J}_c \\ \hline \alpha(1-\beta)\boldsymbol{J}_f & (1-\alpha)\boldsymbol{I} + \alpha\beta\boldsymbol{J}_f\boldsymbol{J}_c \\ \hline \end{array}$$

where:

$\boldsymbol{J}_c = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$ is the Jacobian of the cost functions at point $\boldsymbol{f}$: if it is definite positive the cost functions are strictly increasing;

$\boldsymbol{J}_f = \boldsymbol{Jac}[\boldsymbol{f}_{SUN}(\boldsymbol{c})] = \sum_i d_i \Delta_i \boldsymbol{Jac}[\boldsymbol{p}_i(-\Delta_i^T\boldsymbol{c})]\Delta_i^T$ is the Jacobian of the stochastic uncongested network assignment function, it is symmetric and semi-definite negative under the assumptions that guarantee the monotonicity of the SUN assignment function (see Section 5.3.1).

The elements of the Jacobian $\boldsymbol{Jac}[\psi(\boldsymbol{f}, \boldsymbol{c})]$ depend on the parameters α and β, which specify the choice and utility updating filters respectively. Therefore the values of these parameters affect the stability of a fixed-point.

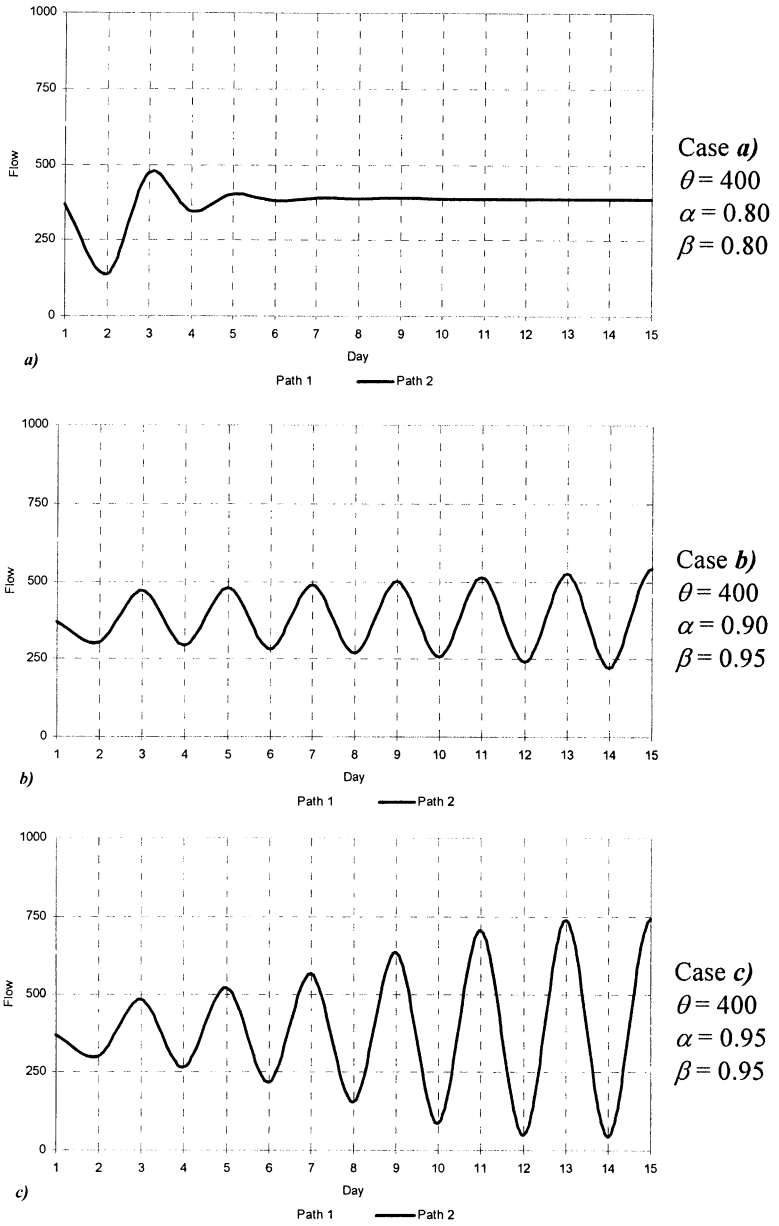


Fig. 5.8.4 Evolution of a deterministic process model for the system in Fig. 5.4.2. (Parameter θ has a different value for highlighting the evolution over time).

In the special case in which $\alpha = \beta = 1$, the Jacobian becomes:

$$\boldsymbol{Jac}[\psi(\boldsymbol{f},\boldsymbol{c})] = \begin{array}{|c|c|}\hline \boldsymbol{0} & \boldsymbol{J_c} \\ \hline \boldsymbol{0} & \boldsymbol{J_f J_c} \\ \hline \end{array}$$

and the eigenvalues are given by:

$$\lambda_j = \gamma_j \quad \forall l$$
$$\lambda_{n+l} = 0 \quad \forall l$$

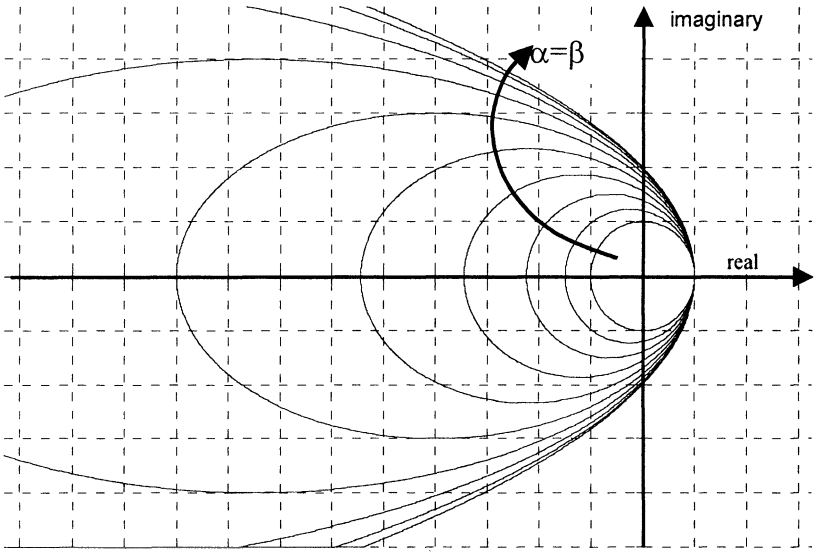
where γ_j is one of the n_L eigenvalues of the matrix $\boldsymbol{J_f J_c}$.

The elements[45] of the matrix $\boldsymbol{J_f J_c}$, and therefore its eigenvalues, depend on the parameters of the system such as the demand flows, the link capacities, the random residuals variance, etc.

In the more general case, if $\alpha \in [0,1]$ and/or $\beta \in [0,1]$, for each of the n_L eigenvalues of the matrix $\boldsymbol{J_f J_c}$, two eigenvalues λ_l and λ_{nL+l} of the Jacobian $\boldsymbol{Jac}[\psi(\boldsymbol{f}, \boldsymbol{c})]$ can be defined as a function of the parameters α and β. The stability condition can be rewritten as a function of the n_L eigenvalues γ_l; it is now represented by an ellipse on the complex plane which must contain the eigenvalues γ_l. In other words, if the system parameters are such that the points representing the n_L eigenvalues γ_l are contained in the ellipse, the fixed-point, or the system's equilibrium state, is stable. This ellipse, whose semi-axes depend only on the parameters α and β, is symmetrical with respect to the real axis and intersects it at two points (see Fig. 5.8.5).

In general, an increase in demand flows and/or a decrease in link capacities and/or a reduction in the variance of the random residuals tends to move the eigenvalues γ_l outside the stability region, while an increase in the parameters α and β tends to reduce the area of the region. Note that whatever the values of α and β, the ellipse is to the left of the point on the real axis with coordinates $\gamma_R = 1$, $\gamma_I = 0$. Therefore, if all the eigenvalues γ_l have a real part less than one, $\gamma_{R,l} < 1$, there are always sufficiently small values of the parameters α and β defining an ellipse sufficiently large to include all the eigenvalues γ_l. In this case, the stability of the fixed-point would be ensured. If the parameters α and β do not satisfy this condition, the fixed point is not stable even if it is unique, and according to results of the theory of non-linear dynamic systems, the system may converge towards quasi-periodic or a-periodic attractors. Vice versa, if some eigenvalues have a real part greater than one, $\gamma_{R,k} \geq 1$, the fixed-point is not stable for any values of α and β; yet there may be other (stable) fixed points.

In the system described in Fig. 5.8.4, for example, (for given path choice and the supply models), as the parameters α and β increase, the system evolves towards attractors other than the fixed point, which becomes unstable. This effect is shown in Fig. 5.8.5.



α=β=1 ; 0.9 ; 0.8 ; 0.7 ; 0.6 ; 0.5 ; 0.4 ; 0.3 ; 0.2 ; 0.1

Fig. 5.8.5 – Stability regions of a fixed point state for α=β.

It is interesting to analyze the relationship between the above considerations and the stochastic equilibrium uniqueness conditions described in section 5.4.1. In particular, it was shown that, under the assumption that the SUN assignment function is monotone with symmetric negative semi-definite Jacobian as described in section 5.3.1, if the cost functions have a positive definite Jacobian $\boldsymbol{J}_c$ (strictly increasing) the stochastic equilibrium is unique. In this case, it can be shown that the eigenvalues γ_l of the matrix $\boldsymbol{J}_f\boldsymbol{J}_c$ always have a non-positive real part, $\gamma_{R,l} \leq 0$. In accordance with the previous considerations, this excludes the possibility of multiple fixed-point states, and therefore of multiple equilibria. Also, it is interesting to note that if the Jacobian of cost functions is symmetric, each of the eigenvalues γ_l of the matrix $\boldsymbol{J}_f\boldsymbol{J}_c$ is real (and non-positive).

A deterministic process model can also be used as an algorithm to find fixed-point attractors, i.e. stochastic equilibrium states. In this case, the model can be defined as a *dynamic process algorithm*, the parameters α and β have no behavioral interpretation and are chosen to guarantee the convergence of the algorithm (i.e. the stability of the fixed point)[46].

5.8.3. Stochastic Process models

Stochastic process models derive from the assumption that path flows (and predicted utilities) on day t, ($\boldsymbol{V}^t$, $\boldsymbol{h}^t$), are random vectors. These models allow one to obtain a statistical description of the system states and to model explicitly some relevant phenomena such as the randomness of link and path performance. The state of the system on the day t, ($\boldsymbol{V}^t$, $\boldsymbol{h}^t$), can therefore be interpreted as a realization of random vectors, ($\boldsymbol{W}^t$, $\boldsymbol{X}^t$). Expressions (5.8.8) and (5.8.9) define the expected values of $\boldsymbol{W}^t$ and $\boldsymbol{X}^t$ as a function of the state ($\boldsymbol{V}^{t-1}$, $\boldsymbol{h}^{t-1}$) on the previous day t-1 and the vector of actual path costs expressed by the random vector $\boldsymbol{G}^{t-1}$:

$$\begin{aligned} &\boldsymbol{V}^t \leftarrow \boldsymbol{W}^t \\ &\text{with } E[\boldsymbol{W}^t] = \boldsymbol{V}(\boldsymbol{V}^{t-1}, \boldsymbol{G}^{t-1}) \end{aligned} \tag{5.8.20}$$

$$\begin{aligned} &\boldsymbol{h}^t \leftarrow \boldsymbol{H}^t \\ &\text{with } E[\boldsymbol{H}^t] = \boldsymbol{R}(\boldsymbol{V}^t)\, \boldsymbol{h}^{t-1} \end{aligned} \tag{5.8.21}$$

Equation (5.8.20) expresses the randomness of the average perceived utilities across the users on the day t, $\boldsymbol{V}^t$. The expected value $E[\boldsymbol{W}^t]$ depends on the actual value of the average path cost on day t-1. The randomness of path costs $\boldsymbol{G}^{t-1}$ might be due to several factors. One of the most important is the randomness of the link costs which, for a given value of h^{t-1}, might take on values c^{t-1} different from the average values, $c(\Delta\, \boldsymbol{h}^{t-1})$. In this case, the link costs c^{t-1} can be modeled as the realization of a random vector $\boldsymbol{C}^t$, and path costs are a linear transformation of $\boldsymbol{C}^t$:

$$\begin{aligned} &c^{t-1} \leftarrow \boldsymbol{C}^{t-1} && \boldsymbol{g}^{t-1} \leftarrow \boldsymbol{G}^{t-1} = \Delta^T \boldsymbol{C}^{t-1} \\ \text{with} \quad &E[\boldsymbol{C}^{t-1}] = c(\Delta \boldsymbol{h}^{t-1}) && E[\boldsymbol{G}^{t-1}] = \Delta^T c(\Delta \boldsymbol{h}^{t-1}) \end{aligned}$$

The randomness of the path flow vector, $\boldsymbol{h}^t$, is dependent on the unpredictability of the users' path choices, whose average value is expressed by the demand model. It is usually assumed that the path flow vector on the day t, $\boldsymbol{H}_{od}^t$, for each pair od is a multinomial random variable; user choices are independent of one another and are made with probabilities given by the demand model $\boldsymbol{R}(\boldsymbol{V}^t)$ as a function of the average perceived utilities across users.

The stochastic process (5.8.20-21) is a *discrete time, homogeneous Markov process*. It is *Markovian* since the state on the day t depends only on the state on the previous day t-1. It is *homogeneous* since the cost and network assignment functions and the cost and choice adjustment parameters are independent of the day. It is *discrete time* since the evolution over time is described by the (integer) index of the day.

Given an initial state ($\boldsymbol{h}^o$, $\boldsymbol{V}^o$), a model of this kind theoretically allows the determination for each day t the probability that the system is in state ($\boldsymbol{h}^1$, $\boldsymbol{V}^1$) belonging to the state space. The probability function, $\phi^t(\boldsymbol{h}, \boldsymbol{V})$, is recursively

defined as the probability that the system is in state $(\boldsymbol{h}^t, \boldsymbol{V}^t)$ on day t conditional on being in state on the previous day t-1, $(\boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1})$:

$$\phi^t(\boldsymbol{h}, \boldsymbol{V}) = Pr[\boldsymbol{h}^t = \boldsymbol{h}, \boldsymbol{V}^t = \boldsymbol{V} / \boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}]$$

Under the assumptions, path flows, $\boldsymbol{h}^t$, are a realization of a discrete random vector, while the predicted utilities, $\boldsymbol{V}^t$, are generally a realization of a continuous random vector. Thus the function $\phi^t(\boldsymbol{h},\boldsymbol{V})$ must be considered a joint probability function with respect to $\boldsymbol{h}$, and a joint probability density function with respect to $\boldsymbol{V}$. In applications to transportation systems, it is often interesting to know the probabilities of path flows (and therefore of link flows). The marginal probability function $\pi^t(\boldsymbol{h})$ of the path flows $\boldsymbol{h}$ on day t is given by:

$$\pi^t(\boldsymbol{h}) = Pr[\boldsymbol{h}^t = \boldsymbol{h} / \boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}] = \int_V \phi^t(\boldsymbol{h}, \boldsymbol{V})\, dV$$

According to the theory of stochastic processes, an *ergodic set* is a minimal subset of the state space such that there is a null probability that the system transitions to a state outside it starting from a state inside it. An ergodic set is minimal in the sense that it does not properly contain ergodic subsets. To each ergodic set is associated a probability function expressing the probability that the system is in a state belonging to the set as $t \to \infty$, known as *stationary probability distribution*:

$$\pi^*(\boldsymbol{h}) = lim_{t\to\infty}\ \pi^t(\boldsymbol{h})$$

Only the states belonging to the ergodic set have a non-null stationary probability. A stochastic process is called *stationary* or *ergodic* if it has respectively one and only one stationary probability distribution $\pi^*(\boldsymbol{h})$. For the specific case discussed here, this stationary probability distribution is $\pi^*(\boldsymbol{h}, \boldsymbol{V})$[47].

A stochastic ergodic process is said to be *regular* if its probability distribution converges towards the unique stationary probability distribution, regardless of the initial state (or its distribution). In this case, a unique (stationary) probability distribution of the system states can be associated with each system specification independently of the initial state. The stationary probabilities $\pi^*(\boldsymbol{h})$, one for each vector $\boldsymbol{h}$ belonging to the ergotic set, can be interpreted as the probabilities of observing the system in any period of observation t sufficiently far from the initial one in the state corresponding to the path flows vector $\boldsymbol{h}$. All the relevant statistics (average, variances, etc.) can be calculated with a single (pseudo-) realization of the process, simulated with Monte Carlo techniques. The transient states visited from a given initial state toward a new stationary distribution following modifications in supply and/or demand can also be analyzed. The probability distribution of each day can be estimated by averaging several (pseudo-) realizations of the process for the same "transient day" t.

A special case, often adopted in applications[48], is obtained if the randomness of the vector of average predicted utilities is ignored, i.e. $\boldsymbol{V}^t = \boldsymbol{W}^t$. This is equivalent to

assuming that the costs realized on day t coincide with the average values given by the cost functions:

$$\boldsymbol{V}^t = \boldsymbol{V}(\boldsymbol{V}^{t-1}, \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^{t-1})) \tag{5.8.22}$$

$$\boldsymbol{h}^t \leftarrow \boldsymbol{H}^t \quad \text{with } E[\boldsymbol{H}^t] = \boldsymbol{R}(\boldsymbol{V}^t)\,\boldsymbol{h}^{t-1} \tag{5.8.23}$$

In this case, the marginal probability $\pi^t(\boldsymbol{h})$, i.e. the probability on day t, of path flows $\boldsymbol{h}$ is given by:

$$\pi^t(\boldsymbol{h}) = Pr[\boldsymbol{h}^t = \boldsymbol{h} \,/\, \boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}] = Pr[\boldsymbol{h}^t = \boldsymbol{h}, \boldsymbol{V}^t = \boldsymbol{V}(\Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^{t-1}), \boldsymbol{V}^{t-1}) \,/\, \boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}]$$

(An example of the more complicated case of a stochastic process with random costs and thus random expected utilities will be presented in section 6.5.2).

It can be demonstrated that the regularity of stochastic processes defined by equations (5.8.22-23) is ensured given the rather general assumptions that the network is connected and that the cost functions and the SUN assignment function are continous. In this case, therefore, a unique probability distribution of path and link flows can be associated with each demand and supply specification, independently of the initial state, and all relevant statistics can be calculated with a single (pseudo-) realization of the process, simulated by Monte Carlo techniques.

According to the law of large numbers, as demand flows increase the evolution of the system described by a stochastic process better approximates the evolution of the corresponding deterministic process model. In this case, the expected values of the path and link flows resulting from a stochastic process model can be approximated well by a corresponding deterministic process, simulating the evolution of average values (process of the averages). From this point of view, stochastic process models seem more suitable for disaggregate, detailed analyses, while deterministic processes are best suited for the simulation of average evolution at an aggregate level and equilibrium stability analyses.

An example of a stochastic process can be obtained by applying exponential filters for the specification of utility and choice updating models:

$$\boldsymbol{V}^t = -\beta \Delta^T \boldsymbol{c}(\Delta \boldsymbol{h}^{t-1}) + (1-\beta)\boldsymbol{V}^{t-1} \tag{5.8.24}$$

$$\boldsymbol{h}^t \leftarrow \boldsymbol{H}^t \quad \text{with } E[\boldsymbol{H}^t] = \alpha\, \boldsymbol{P}(\boldsymbol{V}^t)\boldsymbol{d} + (1-\alpha)\boldsymbol{h}^{t-1} \tag{5.8.25}$$

Similarly, in terms of predicted link flows and costs, we have:

$$\boldsymbol{x}^t = \beta \boldsymbol{c}(\boldsymbol{f}^{t-1}) + (1-\beta)\boldsymbol{x}^{t-1} \tag{5.8.26}$$

$$\boldsymbol{f}^t \leftarrow \boldsymbol{F}^t \quad \text{with } E[\boldsymbol{F}^t] = \alpha \boldsymbol{f}_{SUN}(\boldsymbol{x}^t) + (1-\alpha)\boldsymbol{f}^{t-1} \tag{5.8.27}$$

Another particular stochastic process model is a *renewal process*. These processes are such that the state on day t is a realization of a probability distribution independent from previous days. In this case, the Markovian property of the system expressed by eqns. 5.8.20 and 5.8.21 does not hold. This condition can be formally expressed as:

$$\phi^t(\boldsymbol{h}, \boldsymbol{V}) = Pr[\boldsymbol{h}^t = \boldsymbol{h}, \boldsymbol{V}^t = \boldsymbol{V} / \boldsymbol{h}^{t-1}, \boldsymbol{V}^{t-1}] = Pr[\boldsymbol{h}^t = \boldsymbol{h}, \boldsymbol{V}^t = \boldsymbol{V}]$$

If the joint probability function $\phi^t(\boldsymbol{h}, \boldsymbol{V})$ is constant for each t, the renewal process is stationary. Under these assumptions, renewal process models can simulate systems for which the expected (predicted) utilities of users are independent of the actual costs incurred on previous days (e.g. based on long-term averages or on uncongested values) and there are no habit effects (e.g. $\alpha = 1$ in models 5.8.25 and 5.8.27). An example of a renewal process model in the case of a stochastic supply model with random costs can be expressed by the following equations:

$$\boldsymbol{V}^t = \boldsymbol{W}^t \text{ (MVN variable)}$$
$$\text{with } E[\boldsymbol{W}^t] = -\boldsymbol{\Delta}^T \boldsymbol{c}_o = -\boldsymbol{g}_o$$

$$\boldsymbol{h}^t \leftarrow \boldsymbol{H}^t \text{ (multinomial variable)}$$
$$\text{with } E[\boldsymbol{H}^t] = \boldsymbol{P}(\boldsymbol{V}^t)\boldsymbol{d}$$

A renewal process model will be specified for simulating within-day dynamic irregular transit systems in section 6.5.2.

Finally, it should be noted that the regularity of a stochastic process is a weaker property of the existence, uniqueness and stability of a fixed-point of the corresponding deterministic process. In other words, the stability of a system, in the engineering sense, requires not only the existence of a unique stationary distribution towards which the system state distribution converges, but also that the stationary distribution be unimodal, that is closely spread with respect to a central point.

5.9. Synthesis and application issues of assignment models.

The assignment models described in the previous sections are summarized, using the notation introduced in this chapter, in Fig. 5.9.1 where different models for User Equilibrium assignment are compared, and in Fig. 5.9.2 where "basic" equilibrium models are compared with dynamic process assignment models.

In general, in the case of uncongested networks and rigid demand the assignment model defines a relationship between link flows (output) and link costs and demand flows (inputs). This relationship is defined by UN assignment maps. In the case of congested networks and/or elastic demand, the assignment relationship includes link cost functions and/or demand functions (inputs); this relationship is defined implicitly by equilibrium assignment or dynamic process models. Solution

algorithms for several models are described in Chapter 7, while the methods for estimating the demand flows and the specification and calibration of the demand functions are described in Chapter 8. Cost functions described in Chapter 2 can be calibrated with simple regression analyses on experimental observations when the models reported in the literature are not satisfactory.

The wide range of assignment models described can be used in several contexts and for different classes of application, as will be briefly discussed in the following subsections. The reader should recall that all of the models have been formulated under the assumption of within-day stationarity and therefore implicitly without major over-saturation phenomena that cannot be analyzed in a static context[49]. For this reason, if an assignment model predicts link flows above link capacities for some links, the results of the model can be used as indicators of the critical points of the network, but should no longer be interpreted as estimations of the state of the system. In this case, within-day dynamic assignment models described the following in Chapter 6 should be adopted.

Assignment models as estimators of the present state (monitoring) of the transportation system. In this application, the assignment model receives as inputs the present network and O-D demand flows in order to estimate several quantities which a *would be too costly and complicated to measure directly.* Typically the relevant variables are the flows using different supply elements (road sections, intersection turning movements, lines of public transport services, motorway barriers) represented by links in the network model, the congestion levels of these elements (usually expressed by flow/capacity ratios or load factors), the performance attributes (travel times, monetary costs etc.) comprising the generalized cost of links and paths (used as inputs to demand models), and external impacts (emission and concentration of air pollutants, sound noise pressure, fuel consumption, traffic revenues, etc). In fact, although costs and impacts have been introduced into supply models, in the case of congested networks they depend on link flows and therefore cannot be calculated without the application of an assignment model and its estimated flows. The results of assignment models can complement direct surveys such as flow counts on some links or travel-times measure on some paths, since the latter do not usually cover all the elements of the system. The network variables listed can be used both in project design (identification of critical points, analyses of supply inefficiencies, levels of accessibility, etc.) and in monitoring the effects of planned actions, as will be seen in Chapter 10. For this type of application, rigid (present) demand assignment models can be used.

Assignment models for simulating the effects of modifications to the transportation system. In this application, assignment models are used to estimate the changes in relevant network variables due to changes in supply and/or demand. As will be seen in Chapter 10, this is the typical application of simulation models as "design tools".

		Single-class single-mode assignment with rigid demand	Assignment with pre-trip en-route path choice	Assignment with elastic demand	Multi-mode assignment	Differentiated congestion multi-class assignment	Undifferentiated congestion multi-class assignment
	section	5.2-.3-.4	5.5	5.6.1-.2	5.6.3	5.7.1	5.7.2
SUPPLY MODEL		$\boldsymbol{g}_{od} = \Delta_{od}^{T}\boldsymbol{c} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od} = \Delta_{od}^{T}\boldsymbol{c}$ $\boldsymbol{g}_{od}^{NA}$ *in* $\boldsymbol{x}_{od}^{NA}$	$\boldsymbol{g}_{od} = \Delta_{od}^{T}\boldsymbol{c} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od,m} = \Delta_{od,m}^{T}\boldsymbol{c} + \boldsymbol{g}_{od,m}^{NA}$	$\boldsymbol{g}_{od,i} = \Delta_{od,i}^{T}\boldsymbol{c}^{i} + \boldsymbol{g}_{od,i}^{NA}$	$\boldsymbol{g}_{od,i} = \gamma_i\Delta_{od,i}^{T}\boldsymbol{c}^{i} + \boldsymbol{g}_{od,i}^{spNA}$
		$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$	$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$	$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$	$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$	$\boldsymbol{c}^{i} = \boldsymbol{c}^{i}(\Sigma_i \boldsymbol{f})$	$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$
		$\boldsymbol{f} = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od}$	$\boldsymbol{f} = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od}$	$\boldsymbol{f} = \Sigma_{od}\Delta_{od}\boldsymbol{h}_{od}$	$\boldsymbol{f} = \Sigma_m\Sigma_{od}\Delta_{od,m}\boldsymbol{h}_{od,m}$	$\boldsymbol{f}^{i} = \Sigma_{od}\Delta_{od,i}\boldsymbol{h}_{od,i}$	$\boldsymbol{f} = \Sigma_i\Sigma_{od}\Delta_{od,i}\boldsymbol{h}_{od,i}$
DEMAND MODEL		$\boldsymbol{V}_{od} = -\boldsymbol{g}_{od} + \boldsymbol{V}^{o}_{od}$	$\boldsymbol{V}_{od} = -\Omega_{od}^{T}\boldsymbol{g}_{od} - \boldsymbol{x}_{od}^{NA} + \boldsymbol{V}^{o}_{od}$	$\boldsymbol{V}_{od} = -\boldsymbol{g}_{od} + \boldsymbol{V}^{o}_{od}$	$\boldsymbol{V}_{od,m} = -\boldsymbol{g}_{od,m} + \boldsymbol{V}^{o}_{od,m}$	$\boldsymbol{V}_{od,i} = -\boldsymbol{g}_{od,i} + \boldsymbol{V}^{o}_{od,i}$	$\boldsymbol{V}_{od,i} = -\boldsymbol{g}_{od,i} + \boldsymbol{V}^{o}_{od,i}$
		$\boldsymbol{h}_{od} = \boldsymbol{d}_{od}\boldsymbol{p}_{od}(\boldsymbol{V}_{od})$	$\boldsymbol{h}_{od} = \boldsymbol{d}_{od}\,\Omega_{od}\,\boldsymbol{q}_{od}(\boldsymbol{V}_{od})$	$\boldsymbol{h}_{od} = \boldsymbol{d}_{od}s(\boldsymbol{V}))\,\boldsymbol{p}_{od}(\boldsymbol{V}_{od})$	$\boldsymbol{h}_{od,m} = \boldsymbol{d}_{od,m}\boldsymbol{p}_{od,m}(\boldsymbol{V}_{od,m})$	$\boldsymbol{h}_{od,i} = \boldsymbol{d}_{od,i}\boldsymbol{p}_{od,i}(\boldsymbol{V}_{od,i})$	$\boldsymbol{h}_{od,i} = \boldsymbol{d}_{od,i}\boldsymbol{p}_{od,i}(\boldsymbol{V}_{od,i})$

Fig. 5.9.1.a – Synopsis of User Equilibrium assignment models

	Single-class single-mode assignment with rigid demand	Assignment with pre-trip en-route path choice	Assignment with elastic demand	Multi-mode assignment	Differentiated congestion multi-class assignment	Undifferentiated congestion multi-class assignment
section	5.2-.3-.4	5.5	5.6.1-.2	5.6.3	5.7.1	5.7.2
UNCONGEST. NETWORK ASSIGNMENT MAP	$f_{UN}(\boldsymbol{c};\boldsymbol{d}) = \Sigma_{od}\, d_{od}\, \Delta_{od} \times \times p_{od}(-\Delta_{od}{}^T \boldsymbol{c} - \boldsymbol{g}_{od}^{NA})$	$f_{UN}(\boldsymbol{c};\boldsymbol{d}) = \Sigma_{od}\, d_{od}\, \Lambda_{od} \times \times q_{od}(-\Lambda_{od}{}^T \boldsymbol{c} - \boldsymbol{x}_{od}^{NA})$	$f_{UN}(\boldsymbol{c}) = \Sigma_{od}\, d_{od}(s(-\Delta^T \boldsymbol{c} - \boldsymbol{g}^{NA})) \Delta_{do} \times \times p_{od}(-\Delta_{od}{}^T \boldsymbol{c} - \boldsymbol{g}_{od}^{NA})$	$f_{UN}(\boldsymbol{c}) = = \Sigma_{m,od}\, d_{od,m}(s(-\Delta^T \boldsymbol{c} - \boldsymbol{g}^{NA})) \Delta_{od,m} \times \times p_{od,m}(-\Delta_{od,m}{}^T \boldsymbol{c} - \boldsymbol{g}_{od,m}^{NA})$	$f_{UN}(\boldsymbol{c}^i;\boldsymbol{d}_i) = \Sigma_{od}\, d_{od,i}\, \Delta_{od,i} \times \times p_{od,i}(-\Delta_{od,i}{}^T \boldsymbol{c}^i - \boldsymbol{g}_{od,i}^{NA})$	$f_{UN}(\boldsymbol{c};\boldsymbol{d},\gamma) = \Sigma_{i,od}\, d_{od,i}\, \Delta_{od,i} \times \times p_{od}(-\gamma_i \Delta_{od,i}{}^T \boldsymbol{c} - \boldsymbol{g}_{od,i}^{SPNA})$
SYSTEM OF LINK-BASED EQUATIONS FOR EQUILIBRIUM	$\boldsymbol{c}^* = c(\boldsymbol{f}^*)$ $\boldsymbol{f}^* = f_{UN}(\boldsymbol{c}^*;\boldsymbol{d})$	$\boldsymbol{c}^* = c(\boldsymbol{f}^*)$ $\boldsymbol{f}^* = f_{UN}(\boldsymbol{c}^*;\boldsymbol{d})$	$\boldsymbol{c}^* = c(\boldsymbol{f}^*)$ $\boldsymbol{f}^* = f_{UN}(\boldsymbol{c}^*)$	$\boldsymbol{c}^* = c(\boldsymbol{f}^*)$ $\boldsymbol{f}^* = f_{UN}(\boldsymbol{c}^*)$	$\boldsymbol{c}^{i*} = c^i(\Sigma_i \boldsymbol{f}^*)$ $\boldsymbol{f}^* = \boldsymbol{f}_{UN}(\boldsymbol{c}^*;\boldsymbol{d})$	$\boldsymbol{c}^* = c(\boldsymbol{f}^*)$ $\boldsymbol{f}^* = f_{UN}(\boldsymbol{c}^*;\boldsymbol{d},\gamma)$
FIXED-POINT MODEL FOR EQUILIBRIUM	$\boldsymbol{f}^* = f_{UN}(c(\boldsymbol{f}^*);\boldsymbol{d})$	$\boldsymbol{f}^* = f_{UN}(c(\boldsymbol{f}^*);\boldsymbol{d})$	$\boldsymbol{f}^* = f_{UN}(c(\boldsymbol{f}^*))$	$\boldsymbol{f}^* = f_{UN}(c(\boldsymbol{f}^*))$	$\boldsymbol{f}^* = f_{UN}(c^i(\Sigma_i \boldsymbol{f}^*);\boldsymbol{d})$	$\boldsymbol{f}^* = f_{UN}(c(\boldsymbol{f}^*);\boldsymbol{d},\gamma)$

Fig. 5.9.1.b – Synopsis of User Equilibrium assignment models

		Single-class single-mode assignment with rigid demand	Deterministic process (general example)	Deterministic process (simple example)	Stochastic process (general example)	Stochastic process (simple example)
	section	5.2-.3-.4	5.8.2	5.8.2	5.8.3	5.8.3
SUPPLY MODEL		$\boldsymbol{g}_{od} = \Delta_{od}^T \boldsymbol{c} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od}^{l} = \Delta_{od}^T \boldsymbol{c}^{l} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od}^{l} = \Delta_{od}^T \boldsymbol{c}^{l} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od}^{l} = \Delta_{od}^T \boldsymbol{c}^{l} + \boldsymbol{g}_{od}^{NA}$	$\boldsymbol{g}_{od}^{l} = \Delta_{od}^T \boldsymbol{c}^{l} + \boldsymbol{g}_{od}^{NA}$
		$\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$	$\boldsymbol{c}^{l} = \boldsymbol{c}(\boldsymbol{f}^{l})$	$\boldsymbol{c}^{l} = \boldsymbol{c}(\boldsymbol{f}^{l})$	$\boldsymbol{c}^{l} = \boldsymbol{c}(\boldsymbol{f}^{l})$	$\boldsymbol{c}^{l} = \boldsymbol{c}(\boldsymbol{f}^{l})$
		$\boldsymbol{f} = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}$	$\boldsymbol{f}^{l} = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}^{l}$	$\boldsymbol{f}^{l} = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}^{l}$	$\boldsymbol{f}^{l} = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}^{l}$	$\boldsymbol{f}^{l} = \Sigma_{od} \Delta_{od} \boldsymbol{h}_{od}^{l}$
DEMAND MODEL		$\boldsymbol{V}_{od} = -\boldsymbol{g}_{od} + \boldsymbol{V}^{o}_{od}$	$\boldsymbol{V}_{od}^{l} = \boldsymbol{V}_{od}(\boldsymbol{g}_{od}^{l-1}, \boldsymbol{V}_{od}^{l-1})$	$\boldsymbol{V}_{od}^{l} = -\beta \boldsymbol{g}_{od}^{l-1} + (1-\beta)\boldsymbol{V}_{od}^{l-1}$	$\boldsymbol{V}_{od}^{l} = \boldsymbol{V}_{od}(\boldsymbol{g}_{od}^{l-1}, \boldsymbol{V}_{od}^{l-1})$	$\boldsymbol{V}_{od}^{l} = -\beta \boldsymbol{g}_{od}^{l-1} + (1-\beta)\boldsymbol{V}_{od}^{l-1}$
		$\boldsymbol{h}_{od} = d_{od} \boldsymbol{p}_{od}(\boldsymbol{V}_{od})$	$\boldsymbol{h}_{od}^{l} = \boldsymbol{R}_{od}(\boldsymbol{V}_{od}^{l}) \boldsymbol{h}_{od}^{l-1}$	$\boldsymbol{h}_{od}^{l} = \alpha\, d_{od} \boldsymbol{p}_{od}(\boldsymbol{V}_{od}^{l}) + (1-\alpha)\, \boldsymbol{h}_{od}^{l-1}$	$\boldsymbol{h}_{od}^{l} \leftarrow \boldsymbol{H}_{od}^{l}$ with $E[\boldsymbol{H}_{od}^{l}] = \boldsymbol{R}_{od}(\boldsymbol{V}_{od}^{l})\, \boldsymbol{h}_{od}^{l-1}$	$\boldsymbol{h}_{od}^{l} \leftarrow \boldsymbol{H}_{od}^{l}$ with $E[\boldsymbol{H}_{od}^{l}] = \alpha\, d_{od} \boldsymbol{p}_{od}(\boldsymbol{V}_{od}^{l}) + (1-\alpha)\, \boldsymbol{h}_{od}^{l-1}$

Fig. 5.9.2.a – Synopsis of User Equilibrium and Dynamic Process assignment models (systematic utility is not considered a random variabile for simplicity's sake)

	Single-class single-mode assignment with rigid demand	Deterministic process (general example)	Deterministic process (simple example)	Stochastic process (general example)	Stochastic process (simple example)
section	5.2-.3-.4	5.8.2	5.8.2	5.8.3	5.8.3
UNCONGEST. NETWORK ASSIGNMENT MAP	$f_{UN}(c; d) = \sum_{od} d_{od} \Delta_{od} \times p_{od}(-\Delta_{od}^T c - g_{od}^{NA})$	Not available	$f_{UN}(c; d) = \sum_{od} d_{od} \Delta_{od} \times p_{od}(-\Delta_{od}^T c - g_{od}^{NA})$	Not available	$f_{UN}(c; d) = \sum_{od} d_{od} \Delta_{od} \times p_{od}(-\Delta_{od}^T c - g_{od}^{NA})$
SYSTEM OF LINK-BASED EQUATIONS	$c^* = c(f^*)$ $f^* = f_{UN}(c^*; d)$	Not available	$c^i = \beta\, c(f^{i-1}) + (1-\beta) c^{i-1}$ $f^i = \alpha\, f_{UN}(x^i) + (1-\alpha) f^{i-1}$	Not available	$x^i = \beta\, c(f^{i-1}) + (1-\beta) x^{i-1}$ $f \leftarrow F$ with $E[h^i] = \alpha\, f_{UN}(x^i) + (1-\alpha) f^{i-1}$

Fig. 5.9.2.b – Synopsis of User Equilibrium and Dynamic Process assignment models (systematic utility is not considered a random variabile for simplicity's sake)

The relevant effects of different actions, or projects, are simulated in order to define the technical elements of the project (design) and/or compare alternative hypotheses (evaluation). In this application, the supply and demand models (or the input variables to demand functions) will correspond to the projects and to the future demand scenarios (see section 8.8). If the project network is congested, elastic demand models should be adopted at least for the demand dimensions affected by the planned actions. Different assignment models can be adopted for the design and evaluation phases. Computationally efficient models such as DUE are often used for design, either through supply design models described in Chapter 9 or, through successive trials since several runs are usually required at this stage. Assignment models used to provide measures that allow the comparison of alternative projects should be able to simulate flows and other indicators as accurately as possible, even if at the cost of a greater computation effort.

Assignment models for the estimation of transportation demand. Assignment models are used more and more often for the estimation of O-D demand flows and/or for the calibration of demand models. This type of application, which will be dealt with at length in sections 8.5 and 8.6, "inverts" the usual role of assignment models. When the assignment models are used in this way, they provide relationships connecting present (unknown) O-D flows to the traffic flows measured on some network links, rather than predicting traffic flows from known demand flows. For theoretical reasons regarding the uniqueness of path choice probabilities and flows, it is preferable to use probabilistic (stochastic) assignment models rather than deterministic ones for this purpose.

Interpretation of results and calibration. Regardless of the application, assignment models should be seen as a simplified representations of real, complex phenomena. Thus, the link flows resulting from any assignment model[50] should be correctly indicated with $\boldsymbol{f}^{SIM}$ stressing the fact that they are only estimates of the expected value of the flows $\boldsymbol{f}$ occurring in the real transportation system. The relation between actual flows and the flows resulting from an assignment model can be formally expressed as:

$$\boldsymbol{f} = \boldsymbol{f}^{SIM} + \boldsymbol{\varepsilon}^{SIM} = \boldsymbol{\Delta P}^{SIM}\,\boldsymbol{d} + \boldsymbol{\varepsilon}^{SIM} \tag{5.9.1}$$

The matrix $\boldsymbol{P}^{SIM}$ represents the path choice factions resulting from the assignment model and it generally differs from the matrix $\boldsymbol{P}$ of the "actual" fractions. The vector $\boldsymbol{\varepsilon}^{SIM}$ represents the deviations between actual flows and the flows resulting from the assignment of the demand $\boldsymbol{d}$. These residuals derive from the simplifying assumptions adopted in the system definition (delimitation of study area and zoning) and in the specification of supply, path choice, and supply-demand interaction models as well as in the estimations of the average demand flow $\boldsymbol{d}$. Different assumptions will produce different flows $\boldsymbol{f}^{SIM}$ and residuals $\boldsymbol{\varepsilon}^{SIM}$. This point will be dealt with in greater detail in Section 8.5. For now, note that even if the actual

average demand flows were assigned to the network, all other error sources would produce assignment errors ε^{SIM}.

Assignment models, like all of the mathematical models described in this volume, should be calibrated. The specification of the model and its parameters should reproduce as closely as possible the available data on the state of the system, i.e. minimize the assignment errors ε^{SIM} in expression (5.9.1). However assignment models are affected by several assumptions and parameters since they include all of the assumptions and the parameters of demand and supply models described in this volume. For this reason, a calibration procedure formally derived from the theory of statistical interference has not been proposed. Some partial procedures aimed at selecting assumptions and parameters specific to the assignment model have been applied in a limited number of cases. These usually assume that the supply model and demand functions or O-D flows have been calibrated separately, and focus on the choice of the supply-demand interaction model and the specification and calibration of path choice models.

With respect to *the choice of the supply-demand interaction model*, some experimental evidence indicates that the more realistic the underlying assumptions, the smaller the variance of the deviations ε^{SIM}. For example, for given network and demand flows, both stochastic and deterministic equilibrium models estimate link flows closer to the observed ones than those resulting from uncongested network assignment models, probabilistic models are more accurate than deterministic ones for lightly congested or non-uniformly congested networks, and hyperpath assignment models are more precise than path based assignment models for high-frequency and low-regularity public transport systems. Fig. 5.9.4 reports some experimental curves showing the relative standard deviation, Cv, of the assignment errors, ε^{SIM}, obtained with different assignment models for an urban road network against the counted flows. Relative standard deviations were obtained as the ratios between the standard deviation of the errors between computed and assigned flows in a given range of measured flows and the average flow in the range. Unfortunately, despite the very large number of applications to real transportation systems, the literature proposes few systematic comparative analyses of different assignment models based on large data bases, so that general conclusions on the relative merits of the different models in different application contexts cannot be reached.

Specification and calibration of path choice models can be carried out using disaggregate and/or aggregate data. Disaggregate specification and calibration consists of the selection of the functional form and the attributes (specification) and the statistical estimation of the coefficients (calibration) on the basis of the paths chosen by a random sample of users. Methodologies for disaggregate specification and calibration of path choice models are completely analogous to those used for any random utility model and will be described in Chapter 8. In the case of path choice models, however, disaggregate data are not easy to collect and analyze[51]. Thus aggregate specification and calibration techniques are often adopted. These techniques specify and calibrate path choice models by minimizing a measure of

distance, usually the quadratic errors, between simulated flows, f^{SIM}, and the flows counted on some links. Aggregate calibration of path choice models will be considered again more formally in section 8.6.

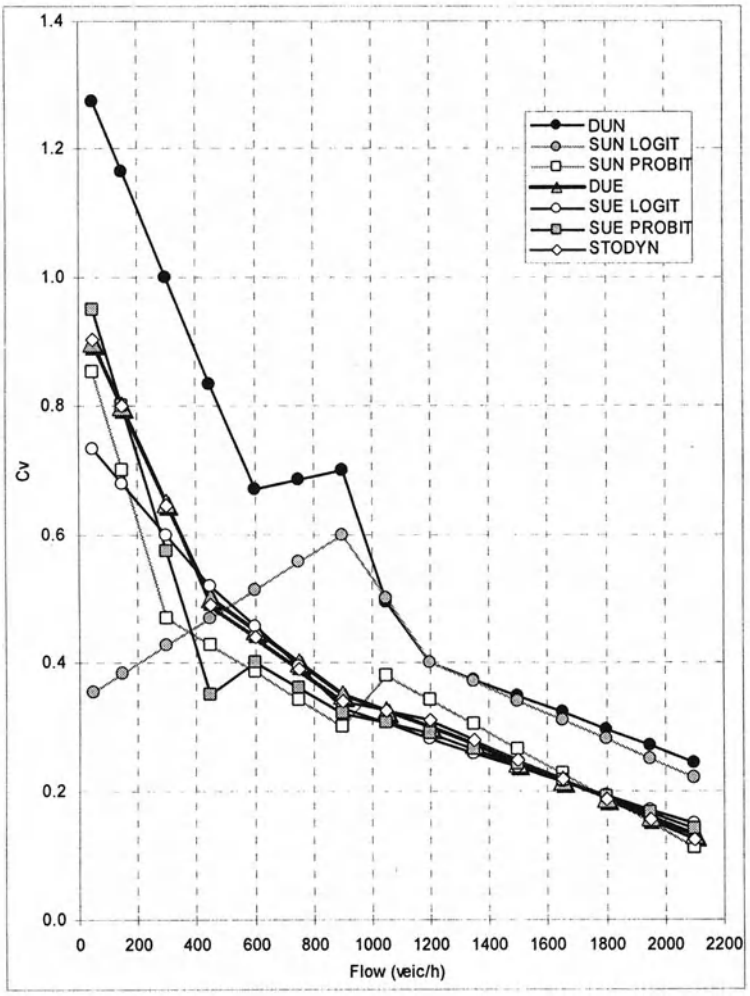


Fig. 5.9.4 Experimental relationships between the coefficient of variation of assignment errors and measured flow.

5.A. Optimization models for stochastic assignment

This appendix describes some optimization models for stochastic assignment, which are equivalent, under some limited assumptions, to the fixed-point models described in the previous sections. Optimization models can also be compared with the deterministic assignment optimization models described earlier. Equivalent optimization models can be used to specificy mathematical programming algorithms for the calculation of stochastic assignment link flows. In some special cases, these algorithms can be reduced to the fixed-point algorithms described in Chapter 7 (e.g. the MSA-FA for stochastic equilibrium), but more generally they are still an open research area. Furthermore, equivalent optimization models for stochastic assignment can be included in bi-level optimization formulations of demand estimation using traffic counts and supply design models. For simplicity of exposition, non-additive path costs are assumed equal to zero.

5.A.1. Stochastic Uncongested Network assignment

For the Logit path choice model with parameter θ independent of link costs, it can be demonstrated that SUN link and path flows are solutions of the following optimization model:

$$(\boldsymbol{f}_{SUN}, \boldsymbol{h}_{SUN}) = argmin\ \Sigma_l c_l f_l + \theta \Sigma_k h_k (ln\ h_k - 1) \qquad (5.A.1a)$$
$$\boldsymbol{f} = \Delta \boldsymbol{h},\ \boldsymbol{h} \in S_h$$

Note that path flows appear explicitly as variables. In terms of the path flows alone, since $\Sigma_l c_l f_l = \Sigma_k h_k g_k$, it follows that:

$$(\boldsymbol{h}_{SUN}) = argmin\ \Sigma_k h_k g_k + \theta \Sigma_k h_k (ln\ h_k - 1) \qquad (5.A.1b)$$
$$\boldsymbol{h} \in S_h$$

It can easily be recognized that the objective functions in models (5.A.1a) and (5.A.1b) are convex if path flows are non-negative.

In both models (5.A.1a) and (5.A.1b), the second term of the objective function goes to zero when the parameter θ goes to zero, i.e. when the variance of path choice random residuals is small. In this case, the path choice model becomes deterministic and both models (5.A.1a) and (5.A.1b) coincide with the optimization model described in Section 5.3.2 for the DUN assignment.

5.A.2. Stochastic User Equilibrium assignment

Similarly to the previous model, for a Logit path choice model with parameter θ independent of link costs, it is possible to demonstrate that stochastic equilibrium link and path flows are solutions of the following optimization model, if cost functions have a symmetric Jacobian:

$$(\boldsymbol{f}^*, \boldsymbol{h}^*) = argmin \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{y})^T \, d\boldsymbol{y} + \theta \Sigma_k h_k (ln\, h_k - 1) \qquad (5.A.2a)$$
$$\boldsymbol{f} = \Delta \boldsymbol{h}, \quad \boldsymbol{h} \in S_h$$

Note that path flows appear explicitly as variables. Since $\boldsymbol{f} = \Delta \boldsymbol{h}$, model (5.A.2a) can be expressed in terms of path flows alone:

$$\boldsymbol{h}^* = argmin \int_0^{\Delta \boldsymbol{h}} \boldsymbol{c}(\boldsymbol{y})^T \, d\boldsymbol{y} + \theta \Sigma_k h_k (ln h_k - 1) \qquad (5.A.2b)$$
$$\boldsymbol{h} \in S_h$$

The objective functions of models (5.A.2a) and (5.4.2b) are (strictly) convex if the path flows are non-negative and the cost functions are (strictly) increasing.

Considering the relationship with the corresponding DUE model, the second term of (5.A.2a) and (5.A.2b) goes to zero as the parameter θ goes to infinity, i.e. as the variance of random residuals gets smaller. In this case, the path choice model becomes deterministic and model (5.A.2a) coincides with the optimization model described in Section 5.4.2 for DUE with symmetric Jacobian cost functions. Fig. 5.A.1 illustrates the equivalent optimization model for SUE Logit assignment for a simple two-link network.

In the case of a general additive path choice model and cost functions with a symmetric Jacobian, it can be shown that equilibrium link flows are a solution of the following constrained optimization model:

$$\boldsymbol{f}^* = argmin_{\boldsymbol{f}} \; \Sigma_{od} d_{od} s_{od}(-\Delta_{od}^T \boldsymbol{c}(\boldsymbol{f})) + \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f} - \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{y})^T \, d\boldsymbol{y} \qquad (5.A.3)$$

where $s_{od} = s_{od}()$ is the path choice EMPU for the pair *od*. Unlike the equivalent optimization model for DUE assignment network, constraints $\boldsymbol{f} \in S_f$ are not needed since they can be proven to be satisfied by all solutions of the model.

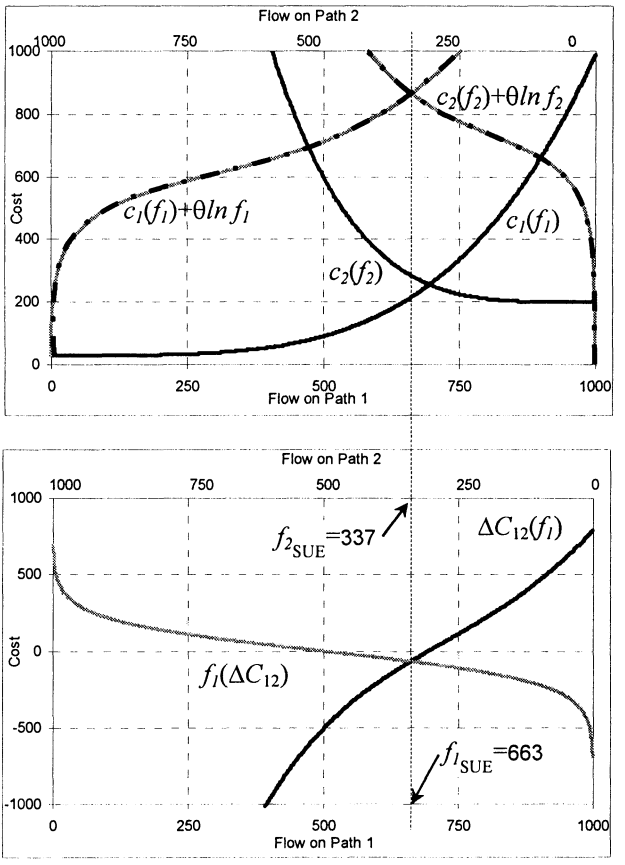


Fig. 5.A.1 – Equivalent optimization model of SUE – Logit assignment .

Reference Notes

The assignment problem has been the subject of extensive research for several decades. Exhaustive analyses of the state of the art of the models (and the algorithms) for uncongested network and user equilibrium assignment are reported in the books by Sheffi (1985), Thomas (1991) and Patriksson (1994), the latter being mainly devoted to deterministic assignment models. For deterministic assignment models, the article by Florian and Hearn (1995) can also be referred to while the state of the art for stochastic assignment models is described in Cantarella and Cascetta (1998).

However the approach to assignment models, or more properly to supply-demand interaction on transportation networks, proposed in this chapter is original. This approach allows through a minimal set of hypotheses and equations to specify consistently uncongested network assignment models as well as fixed-point models and variational inequality models for user equilibrium on congested networks, usually obtained independently from each other. Also the proposed classification of assignment models is an original contribution of this book.

Deterministic User Equilibrium models with separable cost functions and Systems Optimum models were formulated with optimization models analogous to those described in the mid-50s in the pioneering work of Beckman, Mc Guire and Winsten (1956), based from the enunciation of Wardrop (1952) principles. But it was not until the 1970s, with the increasing availability of computing power, that assignment problem received continuous theoretical attention and a number of applications.

The extension of the optimization model to symmetric deterministic equilibrium and the formulation of asymmetric deterministic equilibrium with variational irrequality models, together with existence and uniqueness conditions, are dealt with in the work of Dafermos (1971), (1972), (1980), (1982) and Smith (1979). These articles also describe extensions of DUE models to elastic demand and multi-class assignment. More complex optimization models proposed by various authors for asymmetric deterministic user equilibrium assignment are described and compared by Hearn, Lawphongpanich and Nguyen (1984). Bernstein D. and Smith T. E. (1994) analysed deterministic equilibrium with lower semicontinuous link cost-flow functions.

Extension of deterministic assignment to pre-trip/en-route path choice behavior for transit networks has been proposed Nguyen and Pallottino (1986) and Spiess and Florian (1989). Extensions of DUE assignment models to transit networks and the analysis of its theoretical properties can be found in Nguyen and Pallottino (1988) amd Wu, Florian and Marcotte (1994). Recently, Bouzaiene-Ayari B., Gendreau M., Nguyen S. (1995) and (1997) analysed several approaches to model user behavior at a bus stop within assignment models, including congested waiting times.

Stochastic User Equilibrium (SUE) models were introduced by Daganzo and Sheffi (1977). Optimization models for symmetric SUE were proposed by Fisk

(1980) in the case of the Logit path choice and by Daganzo and Sheffi (1982) in the general case. Bifulco (1993) proposed some extension to simulate parking policies. The extension to pre-trip/en-route path choice behavior for public transportation networks with a Logit-type choice model is dealt with by Nguyen, Pallottino and Gendreau (1993).

Fixed-point models for SUE assignment were introduced by Daganzo (1983), who also analyzed elastic demand assignment (with the hyper-networks approach referred to in Chapter 7) and multi-class assignment. The compact notation and the related reformulation of the optimization problem for SUE models was first adopted by Cascetta (1987). Cantarella (1997) developed a general treatment with fixed-point models of multi-modal/multi-class elastic demand equilibrium assignment also for pre-trip/en-route path choice behavior, including stochastic as well as deterministic user equilibrium.

An analysis of stochastic assignment models with different formulations of random residuals was developed by Mirchandani and Soroush (1987). Nielsen (1997) analysed the advantages and drawbacks of several distributions for link perceived costs. Cantarella and Binetti (2000) described and analysed Gammit path choice models within stochastic equilibrium assignment. Watling (1999) proposed a generalization of SUE models by expressing moments of the distribution of multinomially distributed path flows.

The introduction of capacity constraints in deterministic or stochastic equilibrium models, studied by several authors in the context of a static approach, has been thoroughly analyzed by Ferrari (1997) for deterministic models. Bell (1995) proposed an application for a particular stochastic equilibrium model.

A further line of research relates to equilibrium models in which an (uncongested) cost attribute, such as monetary cost, is distributed among the users, for example following the value of time distribution. These models can be considered an extension of multi-class assignment models to an infinite number of classes in the case of the time value being represented by a continuous random variable. Deterministic equilibrium has been specified with extension of variational inequality models by Leurent (1993, 1995, 1996) and by Marcotte and Zhu (1996), Marcotte et al. (1996), as well as by Dial (1996). The extension of stochastic equilibrium fixed-point models has been dealt with by Cantarella and Binetti (1998).

In recent years dynamic process (non-equilibrium) models for the simulation of supply/demand interaction, have received increasing attention from the scientific community. Initial contributions (Daganzo and Sheffi, 1977, and Horowitz, 1984) analyzed particular models for the study of equilibrium stability; Cascetta (1987, 1989) proposed stochastic process models to represent supply-demand interactions rather than for the analysis of equilibrium configurations. Since then stochastic and deterministic process models have been proposed by various authors among whom Davis and Nihan (1993) and Watling (1996), (1999). An extensive treatment of stochastic and deterministic process models, partily proposed in this chapter, can be found in Cantarella and Cascetta (1995).

Notes

(1) The concept of equilibrium in transportation systems can be compared with the supply-demand equilibrium in classical economics. The analogy, however, is more formal than substantial. As seen in Chapter 2, transportation network cost functions give the cost variation in the use of the system in accordance with variations in the number of system users. Alternatively, in the economic sense, the supply functions relate the service quantity to be produced to the production cost and the sale price of the service. In a given transportation system, and therefore for a given service supply, the equilibrium condition defines the congruence between the demand and the functioning of the supply system, while the equilibrium in a market defines the congruence between the behavior of two "groups": consumers and producers. Furthermore, some special aspects of the transportation system, such as the "network" structure of the supply, make the mathematical treatment of the problem more complex.

(2) In the following sections, the variables corresponding to a path k will usually be identified by the single subscript k, since an od pair is uniquely associated with each path k.

(3) Note that the path cost is measured in units homogeneous with the utility and therefore a change of the measurement units of the attributes which contribute to the definition of the cost does not cause a variation in the value of the systematic utility.

(4) The set S_f of feasible link flows can be also defined without explicitly considering the path flows through a system of linear equations and inequalities with respect to the flow on link l with destination d, say f^d_l. This system expresses the non-negativity of the link flows per destination, the balance between entry and exit flows in each node n, and the total flow on each link l as the sum over all the destinations d of the flows per destination:

$$f_l = \Sigma_d f_l^d \qquad \forall l$$

$$\Sigma_{l \in BS(n)} f_l^d - \Sigma_{l \in FS(n)} f^d = \begin{cases} -d_{od} & \text{if } n = o \\ \sum_o d_{od} & \text{if } n = d \\ 0 & \text{else} \end{cases} \qquad \forall n,d$$

$$f_l^{od} \geq 0 \qquad \forall l,d$$

where

f_l^d is the link flow l with destination d;
$FS(n)$ is the set of exit links from node n, known as the forward star from node n;
$BS(n)$ is the set of entry links into node n, known as the backward star in node n.

The above equations can also be expressed with respect to the flows with a common origin o. They allow one to compare the similarities as well as the differences with the models adopted in other engineering branches for hydraulic or electrical networks.

(5) In the literature these models are sometimes referred to as "network loading" models. In this book that term has been used for a component of the supply model as an alternative to "network flow propagation", it conveys the idea of users moving on the network and inducing link "load" rather than the full demand-supply interaction implied by assignment models. The term "network loading" is also well estabilished with this meaning in the context of within-day dynamic supply models.

(6) The dependence of the flows f on link costs c alone is formally mentioned because it is with respect to these variables that the theoretical properties of the function $f_{SUN}(.)$ will be examined.

[7] By using a stochastic uncongested network assignment, a positive choice probability can also be associated with a non-minimum cost path, equal to the probability that the path has maximum perceived utility (or minimum perceived cost). Because of these considerations, stochastic uncongested network assignment is sometimes indicated as multi-path assignment as compared to the all-or-nothing assignment corresponding to deterministic case.

[8] The model (5.3.6b) can be made formally similar to the model (5.3.5) by considering a further pseudo-link l to which is associated a further row within matrix Δ given by the vector g^{NA} and a cost 1.

[9] Note that by using the definition (5.2.10) in Section 5.2.3, for the feasibility set of link flows S_f an optimization problem is obtained which is known in the literature as linear minimum cost multi-commodity flow.

[10] This assumption can be justified by considering the equilibrium configuration as a state towards which the system evolves (see Section 5.9). From this interpretation, it follows that equilibrium analysis is valid for the analysis of the recurrent congestion conditions of the system; in other words, for those conditions that are systematically brought about by a sequence of periods of reference sufficiently large to guarantee that the system will achieve the state of equilibrium (and remain in it for a sufficient length of time).

[11] If the random residual variance of a path depends on cost, it could happen that as this increases, the corresponding increase in the variance could cause an increase in the path choice probability itself.

[12] In the case of Logit or Probit path choice models, for which a path has a choice probability strictly greater than zero independent of cost, it is possible to demonstrate that the uniqueness of the equilibrium flows is also ensured in the case of cost functions which are not strictly monotone:

$$[c(f') - c(f'')]^T (f' - f'') \geq 0 \qquad \forall f', f'' \in S_f$$

[13] The link cost functions reported in Chapter 2 are all increasing with respect to link flows.

[14] For the deterministic uncongested network assignment map, it is possible to demonstrate properties analogous to those of the stochastic uncongested network assignment function. In particular, the deterministic uncongested network assignment map is semi-continuous, and the set of each link costs vector is non-empty, compact and convex. Furthermore, the map is non-increasing monotone with respect to link costs. These properties permit analysis of the existence and uniqueness of the deterministic user equilibrium flow configurations analogously to the analysis carried out for stochastic user equilibrium flows in Section 5.4.1.

[15] In other words, if and only if the function $c(f)$ has a symmetric Jacobian $\boldsymbol{Jac}[c(f)]$, it can be the gradient of a function $\nabla z(f) = c(f)$, of which Jacobian $\boldsymbol{Jac}[c(f)]$ is the (symmetric) Hessian matrix $\boldsymbol{Hess}[z(f)]$.

[16] Under the same assumptions, a direct demonstration is also possible, even though more complicated, obtained by applying the theory of constrained optimization. Note that the equivalence conditions are stricter than those necessary to define the variational inequality models.

[17] Note that by using the definition (5.2.10) in Section 5.2.3, for the feasibility set of link flows S_f an optimization problem is obtained which is known in the literature as convex minimum cost multi-commodity flow.

[18] Formal models for supply design will be dealt with in section 8.4.

[19] Results reported in the literature indicate that cost functions characterized by a stronger form of monotonicity exclude the occurrence of the Braess' paradox.

[20] Analogously to the content of previous sections, it is assumed that the users belong to a single class, or that they undertake trips for the same purpose and have equal hyperpath choice models. Generalization to multi-class assignment will be dealt with in section 5.7.

[21] In other words, it is assumed that service congestion affects the perceived cost of on-board time, but not waiting time. This excludes the possibility of waiting longer because of congestion since some runs are not available because they are too crowded.

[22] In the case of Logit or Probit path choice models, for which a path has a choice probability strictly greater than zero independent of cost, it can be demonstrated that the uniqueness of the equilibrium flows is also ensured in the case of cost functions which are not strictly monotone:

$$[c(\boldsymbol{f}') - c(\boldsymbol{f}'')]^T (\boldsymbol{f}' - \boldsymbol{f}'') \geq 0 \qquad \forall \boldsymbol{f}', \boldsymbol{f}'' \in S_f$$

[23] Rigid demand assignment models also occurs when demand flows are assumed to be dependent on path cost attributes independent of congestion, or independent of the flows, as for example null-flow or null-distance etc. generalized times or costs.

[24] It is also possible to adapt rigid demand assignment models to deal with elastic demand, by expanding the network model with links which are appropriately defined (hyper-networks). In this way, the choice behavior on other dimensions can be simulated like path choice in a modified network. This approach is difficult to generalize and is not subject to further analysis (see also section 7.6).

[25] A strictly monotone function is invertible, and an invertible and continuous function is strictly monotone (see Appendix A).

[26] It should be noted that is is very difficult to get closed-form expressions for the inverse demand functions, $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{d})$, even in the case of simple demand models. This characteristic considerably limits the application of variational inequality models for elastic demand deterministic equilibrium (but not of fixed-point models). In the case of Logit-type demand models, an optimization model can be adopted as will be seen below.

[27] To this end, note that both the models (5.6.6) and (5.6.7) can be expressed as a variational inequality defined on the set S of the function $\varphi(\boldsymbol{x})$ with respect to a vector $\boldsymbol{x}$: $\varphi(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \geq 0$, $\forall \boldsymbol{x} \in S$. In particular, in the model (5.6.6) the vector $\boldsymbol{x}$ is defined by the path and demand flows vectors, $\boldsymbol{h}$ and $\boldsymbol{d}$, the set S is defined by the product of the set of feasible path and demand flows, S_h and S_d, and the function $\varphi(\boldsymbol{x})$ is defined by the path costs functions and the negative of the inverse demand function, $\boldsymbol{g}(\boldsymbol{h})$ and $-\boldsymbol{Z}(\boldsymbol{d})$. The same holds for the model (5.4.7) expressed in terms of link flows and demand flows.

[28] Note that strict monotonicity is needed, in contrast to stochastic user equilibrium.

[29] In the special case in which each link can be used by one mode only, and the cost on a link depends only on the flows of the corresponding mode, the entire network is separable into independent modal networks.

[30] At the most, each segment can consist of a single user, and in this way disaggregated assignment models are obtained. Models of this type are at present only in the research stage.

[31] Different classes corresponding to the same od pair may have different incidence matrices if they have different available path sets.

[32] It is also possible to specify cost functions for class i depending only on the flow $\boldsymbol{f}$; these models, however, are seldom adopted as they do not correspond to known congested phenomena.

[33] The supply model (5.7.4) can also be interpreted as an application of the general network supply model (5.2.4) in which each "physical link" is represented by several "network links", one for each class.

[34] Note that the two conditions (5.7.9) and (5.7.10) coincide if two flow vectors are considered which differ only in terms of class flows. The same circumstance obviously occurs in the case of a single class of users.

[35] More generally, note that the results of deterministic path choice models are not modified even by a non-linear relationship between systematic utilities and path cost, as long as this relationship is strictly increasing.

[36] For the sake of simplicity, the generic reference period will be identified as a "day". Note that the periods need not be successive. For example, reference can be made only to weekdays or to periods of "fictitious" behavior updating if the aim is not to explicitly simulate the development of the system but only its convergence properties.

[36] A dynamic process assignment model can also be multi-class and applied to different levels of aggregation by considering for each O-D pair homogeneous classes of users, each consisting, in the extreme case, of a single user (completely disaggregated assignment).

[38] Note that the two cost update models, or systematic utility models, correspond to two assumptions which differ in terms of the underlying behavioral mechanism. In the case of the model (5.9.5c), it is assumed that the user remembers and averages path costs on successive days while in the case of the model (5.9.5d), it is assumed that memory is relative to the costs of individual links, which are put together later to obtain the paths values. Based on the assumptions made, the two formulations are equivalent, but they might not be for other costs updating models different from those described and/or in the presence of non-additive path costs.

[39] In some more sophisticated formulations of the choice updating model, it is assumed that the parameter is substituted with a model expressing the probability of reconsidering the choices in variables of the socio-economic and service-level type (difference between expected values and actual values, information, etc.).

[40] The adoption of different formulations for the cost and choice updating models can lead to a different definition of the state of the system. For example, if an average mobile filter on k previous days is adopted to specify the cost updating model, the state of the system on day t is defined by the path flows together the path costs on k previous days .

[41] In other words, N is the number of the components of the vector that describes the state of the system. In the case of the model (5.9.10.), we get $2n_p$ where n_p is the number of paths.

[42] The boundary points between different attractor basins are singular points of behavior (saddle points, for example) which can be ignored in a first analysis in that small variations in the initial state move the development of the system toward the basin of an attractor.

[43] This condition, not generally valid, can be extended to a larger class of cost (but not choice) updating models.

[44] An eigenvalue of a square matrix $\boldsymbol{J}$ is a number λ respecting the condition: $\boldsymbol{J}\omega = \lambda\omega$, with $\omega \neq \mathbf{0}$. The vector ω is known as an eigenvector of the matrix $\boldsymbol{J}$ corresponding to the eigenvalue λ. Eigenvalues are the solutions of the algebraic equation $|\boldsymbol{J} - \lambda\boldsymbol{I}| = 0$, equal in number to the dimensions of the matrix $\boldsymbol{J}$. A real matrix can have real or complex eigenvalues (and eigenvectors) in conjugate pairs, a symmetric matrix can only have real eigenvalues (and eigenvectors).

(45) The elements of the matrix $J_f J_c$, and therefore its eigenvalues a-dimensional and the stability of a fixed point is therefore not influences by the unit of measurement adopted.

(46) For this purpose, compare the algorithm MSA-FA, described in Section 7.4.1, with the model (5.9.16-17), assuming: $\alpha = 1/t$, $\beta = 1$.

(47) A stochastic process can be interpreted as a deterministic process in the space (of infinite dimensions) of the density functions $\pi(\boldsymbol{h})$, whose state on day t is given by $\pi^t(\boldsymbol{h})$. In this interpretation, a, ergodic set is a fixed-point state of the deterministic process. The properties of stationarity, ergodicity and regularity correspond to the existence, uniqueness and (global) stability of this fixed-point state, which is a deterministic process attractor in that it is (globally) stable.

(48) The proposed formulation could easily be extended also to consider the costs as random variables. Note, however, that by adopting a probabilistic path choice model a perceived utility of randomness is introduced which can also be attributed implicitly to the randomness of the attributes which appear in the systematic utility (in this case, the path costs).

(49) On the other hand, as was noted in Section 5.4.1, the introduction of capacity constraints in static models is not convenient or simple to interpret, and within-day dynamic assignment models are still being researched (see Chaper 6).

(50) Completely analogous considerations can be made in relation to the other variables resulting from the assignment model such as link costs, path costs and flows, performances, etc.

(51) In reality, it is often a complex task to determine the path actually followed during a journey. Also, even in the case in which there is a path choice model specified and calibrated on disaggregated data, it is useful to carry out an aggregated recalibration which, from a theoretical point of view, can be seen as a correction of the parameters to compensate for the errors of the aggregation process of disaggregated models.

6 INTRA-PERIOD (WITHIN-DAY) DYNAMIC MODELS*

6.1. Introduction

The mathematical models described in the previous chapters are based on the assumptions of intra-period stationariety. This is equivalent to assuming, as stated in Chapter 1, that all significant variables are constant, at least on average, over successive sub-intervals of a reference period long enough to allow the system to reach stationariety condition. This assumption, although acceptable for many applications, does not allow for the satisfactory simulation of some transportation systems such as heavily congested urban road networks or low frequency scheduled services. In the first case, some important phenomena cannot be reproduced by traditional intra-period static models, including demand peaks, temporary capacity variations, temporary over-saturation of supply elements, and formation and dispersion of queues. In the second case, low-frequency services (e.g. two flights per day) may call into question the assumption of intra-period uniform supply and mixed preventive-adaptive users' choice behavior introduced in the previous chapters. To simulate these aspects, different intra-periodal or within-day dynamic models have recently been developed; these models are usually referred to in the literature as (within-day) Dynamic Traffic Assignment (DTA) models, implying that dynamic assignment models require within-day dynamic demand and supply models.

The variables of intra-periodal dynamic transportation systems are indexed by time, τ, internal to the reference period. Within-day dynamic models, however, should not be seen simply as extensions of static models described in the previous chapters with the addition of a further index τ. In fact, these models require a substantial reformulation of demand and, most importantly, supply models. This chapter covers extensions of static supply, demand and assignment models to the within-day dynamic case. In addition, it will be assumed that demand flows within the reference period are known and modeled as described in Chapter 4 on all dimensions but route and departure time.

Intra-period dynamic models have different formulations and levels of complexity depending on the type of supply system involved. As seen in Chapter 2, transportation services and representative supply models can be divided in two main classes: continuous and scheduled. The first case considers services available at any

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

time and accessible from several points, such as the services offered by individual road modes (car, bicycle etc.). On the other hand scheduled services are available only at certain times and can be accessed only at certain locations (terminals, stations, airports etc.). As will be shown more clearly later in this chapter, extensions of static models to within-day dynamic models are easier for scheduled services since the discrete time framework allows a simpler modeling of supply and network flow propagation or network loading.

In the following of this chapter, section 6.2 discusses within-day dynamic supply, section 6.3 within-day dynamic demand and section 6.4 supply-demand interaction models for continuous service (road) systems extending the results presented in Chapters 2 and 5. Section 6.5 describes within-day dynamic models for scheduled services systems. Throughout the remainder of this chapter, to simplify notation and analysis, a single user class and rigid origin-destination demand flows will be assumed.

6.2. Supply models for continuous service systems

Dynamic models for continuous service networks can be categorized by the representation of user flows: continuous or discrete (see Appendix 2.A).

In *continuous flow* models, users are modeled as "particles" of a mono-dimensional, partly compressible fluid, moving at different rates through the system. On the other hand, *discrete flow* models assume that users are discrete units; these can be packets, e.g. groups of vehicles sharing the same trip, or individual vehicles. Flows are defined as the number of user units moving in a time interval. Continuous flow models represent the "natural" extensions of static supply, demand and assignment models; on the other hand discrete flows are more consistent with the reality and with the computational framework. In the following, supply, demand and demand-supply interaction models will be described for both continuous and discrete flow models.

Within-day dynamic supply models, like static models, express flows and performances of the system (flows on individual links, travel times, generalized costs, etc.) as functions of the path flows and the characteristics of the physical system. Although the components of a dynamic supply model are the same, the relationships between link and path flows and costs are no longer linear as in static models (see equations 5.2.3 and 5.2.1).

Fig. 6.2.1 reproduces the general structure of a transportation supply model similar to that introduced in Chapter 2; the only visible difference is the dependence of the network flow propagation model on link performances.

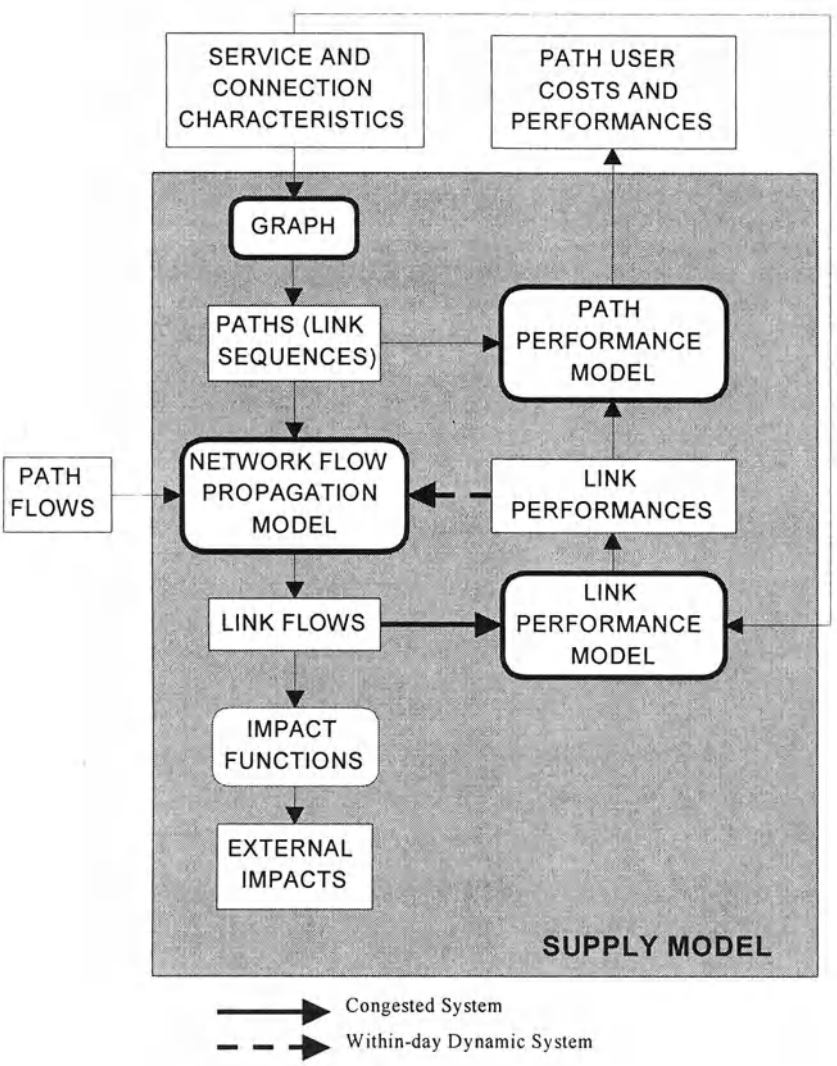


Fig. 6.2.1 Schematic representation of within-day dynamic supply models.

In the first component of the supply model, the different phases of journeys on a continuous supply system are represented with a *graph model*. This model is based on a graph representation of relevant phases of a trip (links) and relevant positions in space and/or time (nodes) analogous to those introduced in Section 2.2.1. For these systems all concepts and notation related to graph models for continuous transportation services such as centroids, paths, incidence matrices etc. extend

directly to within-day dynamic networks. In applications two types of links are typically considered: *running links* representing movement phases, such as the movement along a motorway or an urban road section, and *queuing or waiting links* representing the waiting phenomena at intersections, tollbooths, etc. (see Fig. 6.2.2). As in static models, the same physical system can be represented with different levels of spatial and/or functional detail. The description of these possibilities is outside the scope of this section.

The formulation of general expressions for the other components of a dynamic supply model depends on the basic assumptions on the flow structure, i.e. whether a continuous or discrete approach is followed. In analogy with static models, supply models will be formulated for the case of continuous-time continuous-flow first; successively they will be extended to the discrete case.

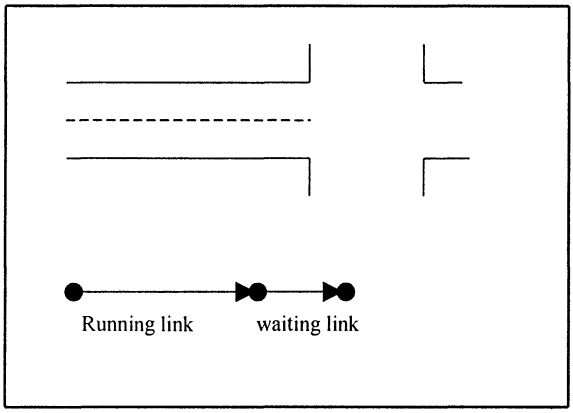


Fig. 6.2.2 Representation of a road intersection with running and waiting links.

6.2.1. Continuous flow supply models

Continuous flow models can be further classified by the representation of space. *Space discrete* (link-based) models are closer to static models: the basic variables influencing link performances such as densities and speeds are defined with respect to links. *Space continuous* models are, on the contrary, based on variables defined for single points in space. This model can be obtained from the macroscopic flow models with continuos flow representation, briefly described in section 2.A.1.3. The solution of these models however requires a discretization in time and space, as noted at the end of this section. The following will focus mainly on link-based models.

6.2.1.1. Variables and consistency conditions

Variables of continuous flow link-based dynamic supply models can be classified into three groups: topological, flow/occupancy, time/cost variables.

Topological variables. The topological features of a journey are modeled through a graph. Let

a be the index of a link of length L_a;
k be the index of a path, made up of a sequence of links $a^k_1, a^k_2, \ldots .a^k_{n_k}$ where a^k_i is the *i-th* link of path k, n_k the number of link of path k;
a^k_{i+1} be the link following a^k_i on path k;
a^k_{i-1} be the link preceding a^k_i on path k.

Flow and occupancy variables. For analytical convenience it will be assumed that all the flow variables are continuous and continuously differentiable functions of time $\tau \geq 0$. Let

$d_{od}(\tau)$ be the origin-destination demand flow at time τ, i.e. the flow rate of users leaving zone (node) o at time τ to zone (node) d;
$\boldsymbol{d}(\tau)$ be the vector of O-D flows at time τ;

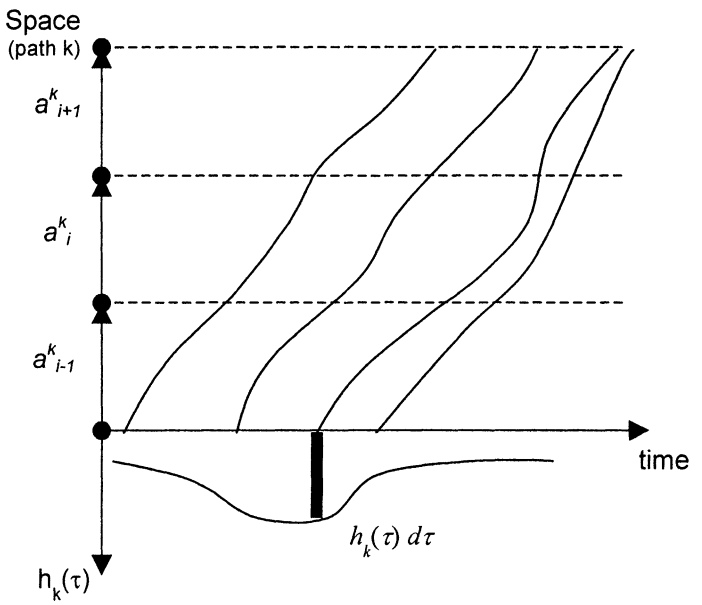


Fig. 6.2.3 Path flow and trajectories in continuous flow models.

$h_k(\tau)$ be the path flow of users who start their journey at time τ and follow path k; under the assumptions made, users are no longer discrete elements but fluid particles leaving on path k at a time density given by $h_k(\tau)$, the time-space trajectories of the particles leaving at each point in time can be traced along the links making up the path. The number of users leaving on path k in the infinitesimal interval $(\tau, \tau+d\tau)$ is equal to $h_k(\tau)d\tau$ (see Fig. 6.2.3);

$\boldsymbol{h}_{od}(\tau)$ be the path flow vector with components given by the flows $h_k(\tau)$ relative to each path connecting od pair, $k \in K_{od}$;

$\boldsymbol{h}(\tau)$ be the total vector of path flows for all the O-D pairs at time τ;

$f^k_{as}(\tau)$ be the user flow following path k and crossing section s of link a at time τ. Unlike the static case, it is not possible to define a generic link flow since flow crossing at the time τ the different sections of a link usually is not constant over the link (see Fig. 6.2.4). Among the sections of a link, entrance (s=0) and exit (s=L_a) are particularly relevant;

$u^k_a(\tau)=f^k_{a,0}(\tau)$ be the flow traveling on path k and entering link a at time τ (*in-flow*), $u^k_a(\tau) \geq 0$;

$w^k_a(\tau)=f^k_{a,L_a}(\tau)$ be the flow traveling on path k leaving link a at time τ (*out-flow*), $w^k_a(\tau) \geq 0$;

$f_{a,s}(\tau)$, $u_a(\tau)$, $w_a(\tau)$ be the total flow crossing section s, entering and leaving link a, at time τ, respectively, and relate to the path-specific variables through the following relationships:

$$f_{a,s}(\tau) = \sum_k f^k_{a,s}(\tau) \tag{6.2.1a}$$

$$u_a(\tau) = \sum_k u^k_a(\tau) \tag{6.2.1b}$$

$$w_a(\tau) = \sum_k w^k_a(\tau) \tag{6.2.1c}$$

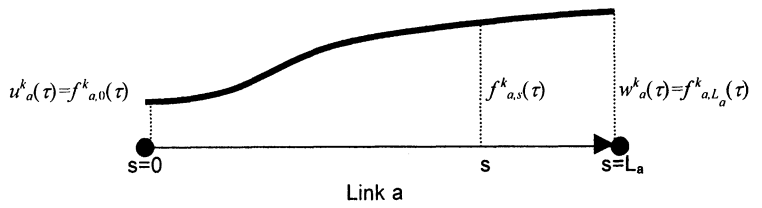


Fig. 6.2.4 Instantaneous flows on link sections.

$U_a(\tau)$ $W_a(\tau)$ be the cumulative in-flow and out-flow on link a at time τ, respectively; they express the total number of users who entered and left the link up to

a given point in time τ. Cumulative flows relate to flow rates through the following equations:

$$U_a(\tau) = \int_0^\tau u_a(t)dt$$

$$W_a(\tau) = \int_0^\tau w_a(t)dt$$

$x_a(\tau)$ be the number of users on link a at time τ or link occupancy;
$k_a(\tau)=x_a(\tau)/L_a$ be the users density on link a at time τ.

Temporal profiles of flow variables must satisfy conservation equations since users flows cannot be created or dispersed at any point of the network except centroid nodes. If no flows are generated and/or absorbed at a node i (i.e. the node i is not a centroid node) flow conservation conditions require that in-flows, out-flows and path flows satisfy the following equations:

$$u^k_{a^k_1}(\tau) = h_k(\tau) \tag{6.2.2a}$$

$$u^k_{a^k_{i+1}}(\tau) = w^k_{a^k_i}(\tau) \tag{6.2.2b}$$

and summing over all paths, equations (6.2.2b) yield:

$$\sum_{a\in FS(i)} u_a(\tau) = \sum_{a\in BS(i)} w_a(\tau) \tag{6.2.2c}$$

Equation (6.2.2c) constrains the total out-flow of the links belonging to its backward star $BS(i)$ to equal the total in-flow on all the links belonging to its forward star $FS(i)$ at any time τ for a node i that is not a centroid. Equations (6.2.2) can be extended easily for centroid nodes distinguishing between path ending and/or starting in node i.

As shown in chapter 2, the link density at time τ can be expressed as a function of (non-negative and integrable) in-flow and out-flow temporal profiles. The differential equation expressing flow conservation on a link is:

$$\frac{dx_a(\tau)}{d\tau} = u_a(\tau) - w_a(\tau) \tag{6.2.3}$$

which, once integrated, leads to the following result:

$$x_a(\tau) = L_a \cdot k_a(\tau) = \int_0^{\tau} u_a(t)dt - \int_0^{\tau} w_a(t)dt = U_a(\tau) - W_a(\tau) \qquad (6.2.4)$$

Equation (6.2.4) expresses the relationship among density and cumulative in-flow and out-flow on a link. The number of users on link a at time τ is the difference between the cumulated in-flow and out-flow at that time, see Fig. 6.2.5.

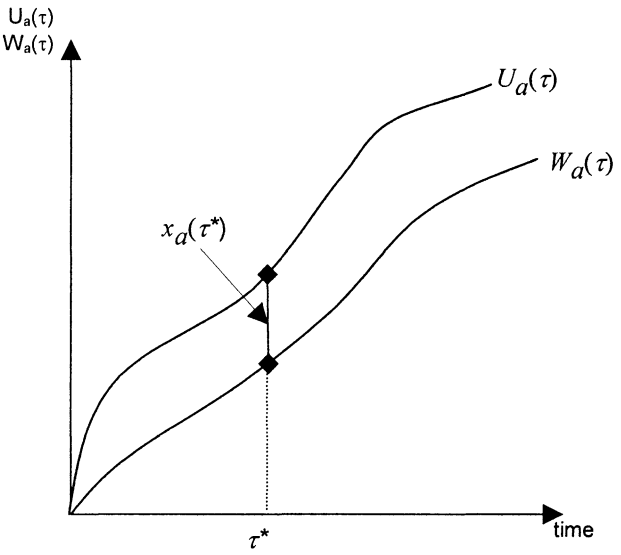


Fig. 6.2.5 Relationship between cumulated in-flow, cumulated out-flow and link occupancy (continuous flows).

Travel time and cost variables. In within-day dynamic supply models the travel time has a dual role. It is a performance variable included as an attribute of the generalized perceived cost, as in within-day static models. On the other hand, it ensures the internal consistency of the relationships between some variables of the model. For this reason, travel time is denoted with a specific variable different from the other performance variables. Moreover, travel times of links and paths may assume different values for different time instants; this may depend on different transportation supply and/or congestion conditions. For this reason, a number of new variables related to travel time must be introduced. It will be assumed that travel times are continuous and continuously differentiable functions of the absolute time τ. Let

$t_a^f(\tau)$ be the forward travel time, i.e. the time to cross link a for a flow particle entering the link at time τ;

$t_a^b(\tau)$ be the backward travel time, i.e. the time to cross the link a for a flow particle leaving the link at time τ;

$t_a^L(\tau)$ be the *leaving-time* function, representing the leaving time of a particle entering link a at time τ;

$t_a^{-1}(\tau)$ be the *inverse travel time* function, representing the entrance time of a particle leaving link a at time τ;

$ec_a(\tau)$ be the generalized extra cost for crossing link a entering at time τ. The generalized extra cost expresses the perceived disutility of link a with the exception of the travel time. It includes other performance variables, e.g. time variable tolls, homogenized in disutility terms.

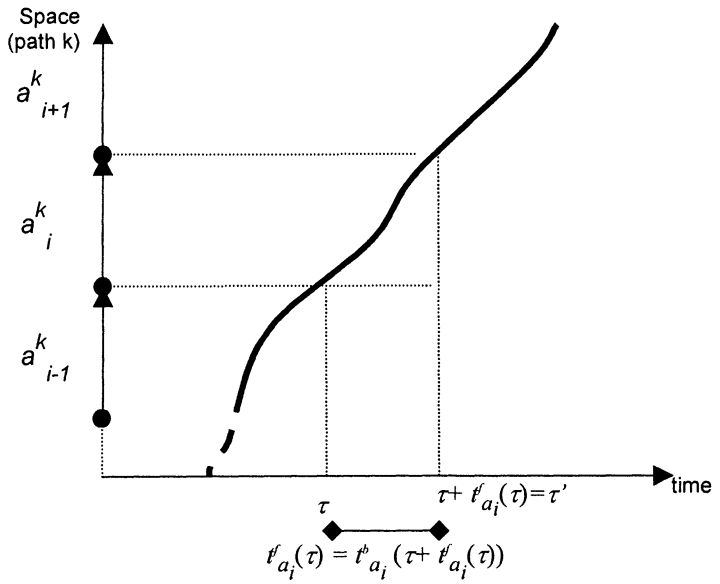


Fig. 6.2.6 Relationship between backward and forward travel times.

Temporal consistency of the model requires that the different travel times satisfy the following relationships (see Fig. 6.2.6):

$$t_a^f(\tau) = t_a^b(\tau + t_a^f(\tau)) \tag{6.2.5a}$$

$$t_a^b(\tau) = t_a^f(\tau - t_a^b(\tau)) \tag{6.2.5b}$$

$$t_a^{-1}(\tau) = \tau - t_a^b(\tau) \tag{6.2.5c}$$

$$t_a^L(\tau) = \tau + t_a^f(\tau) \tag{6.2.5d}$$

Travel times themselves must be consistent. In fact, under the assumption of partly compressible mono-dimensional fluid, travel time functions must be such that a fluid particle entering at time τ'' on link a can never reach, or overtake, another particle that entered the same link at an earlier time $\tau' < \tau''$. If this were the case, it would imply the fluid left between τ' and τ'' be compressed to a zero space (infinite density) or, in the case of overtaking, the mono-dimensionality assumption of the fluid (no turbulence along a link) would be violated. This condition is usually referred to in the literature as strong First-In-First-Out (FIFO) rule, see also Appendix 2.A, and can be stated formally as (see Fig. 6.2.7):

$$\tau' + t_a^f(\tau') < \tau'' + t_a^f(\tau'') \qquad \forall \quad \tau' < \tau'' \tag{6.2.6a}$$

similarly for the backward travel time:

$$\tau' - t_a^b(\tau') < \tau'' - t_a^b(\tau'') \qquad \forall \quad \tau' < \tau'' \tag{6.2.6b}$$

Weak FIFO rule is obtained when strict inequality is substituted by weak inequality within the above conditions. For sake of brevity this topic will not be discussed in the following.

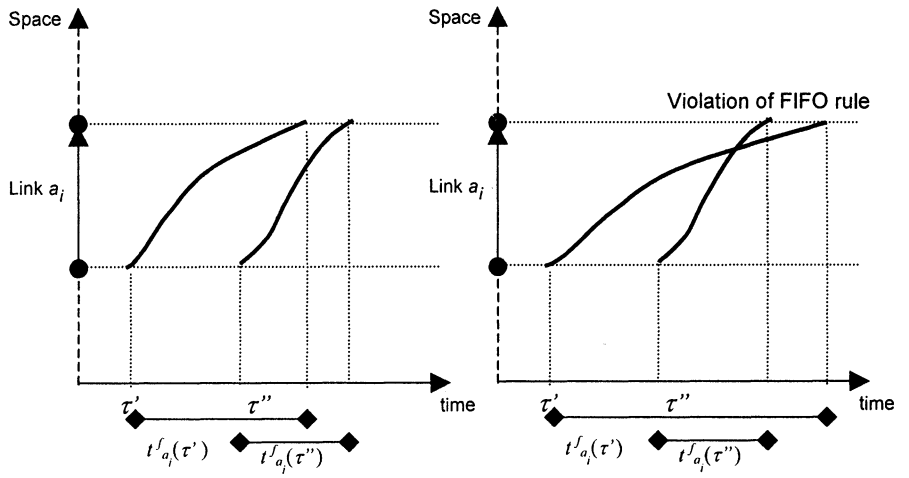


Fig. 6.2.7 Representation of FIFO rule on a link.

Relationships (6.2.6a) and (6.2.6b) imply that $\tau + t_a^f(\tau)$ and $\tau - t_a^b(\tau)$ are well-defined strictly increasing functions of τ, i.e. a single value of exit or entrance time correspond to each value of τ. On the other hand, as said above, without the FIFO rule, two particles could cross the same section at the same time. Thus, a single value of absolute time may correspond to different values of speed and acceleration in the same point in space and the inverse of time function $t^{-1}{}_a$ would be ill-defined.

It can be shown easily that a sufficient condition for a FIFO discipline (6.2.6a) is the following:

$$\frac{dt_a^f(\tau)}{d\tau} > -1 \qquad \forall \tau \tag{6.2.7a}$$

In fact, equation (6.2.7a) can be also stated as:

$$\lim_{\tau'' \to \tau'} \frac{t_a^f(\tau'') - t_a^f(\tau')}{\tau'' - \tau'} > -1$$

which yields

$$\frac{t_a^f(\tau'') - t_a^f(\tau')}{\tau'' - \tau'} > -1$$

thus

$$\frac{t_a^f(\tau'') - t_a^f(\tau')}{\tau'' - \tau'} + \frac{\tau'' - \tau'}{\tau'' - \tau'} > 0$$

and

$$\frac{\tau'' + t_a^f(\tau'') - \tau' - t_a^f(\tau')}{\tau'' - \tau'} > 0 \qquad \forall \tau'' > \tau'$$

which is condition (6.2.6a).

Similarly in terms of backward travel time it can be shown that a sufficient condition for a FIFO discipline (equations 6.2.6b) is:

$$\frac{dt_a^b}{d\tau}(\tau) < +1 \qquad \forall \tau \tag{6.2.7b}$$

The physical interpretation of equations (6.2.7) is that to avoid FIFO rule violation, the travel time cannot decrease more rapidly than the absolute time.

Several equivalent conditions have been proposed to impose a FIFO discipline. One of the most intuitive states that a FIFO discipline exists if and only if the total number of vehicles entering a generic link a by time τ equals the total number of

vehicles exiting after a time interval equal to the forward travel time of link a at any time τ (see Fig. 6.2.8):

$$U_a(\tau) = W_a(\tau + t_a^f(\tau)) \qquad \forall \tau \tag{6.2.8a}$$

similarly in terms of backward travel time:

$$U_a(\tau - t_a^b(\tau)) = W_a(\tau) \qquad \forall \tau \tag{6.2.8b}$$

By deriving the above relationships it follows that:

$$u_a(\tau) = w_a(\tau + t_a^f(\tau)) \cdot \left(1 + \frac{dt_a^f}{d\tau}(\tau)\right) \tag{6.2.9a}$$

$$u_a(\tau - t_a^b(\tau)) \cdot \left(1 - \frac{dt_a^b}{d\tau}(\tau)\right) = w_a(\tau) \tag{6.2.9b}$$

Equations (6.2.9) express the relationship between link in-flows and out-flows in a dynamic context. From equation (6.2.9a) the in-flow on a link at time τ is equal to out-flow at the corresponding exit time (i.e. the absolute time after the link forward travel time) multiplied by a factor larger than one (i.e. the out-flow is less than the in-flow) if the flow on the link is slowing down ($dt/d\tau > 0$) and vice versa.

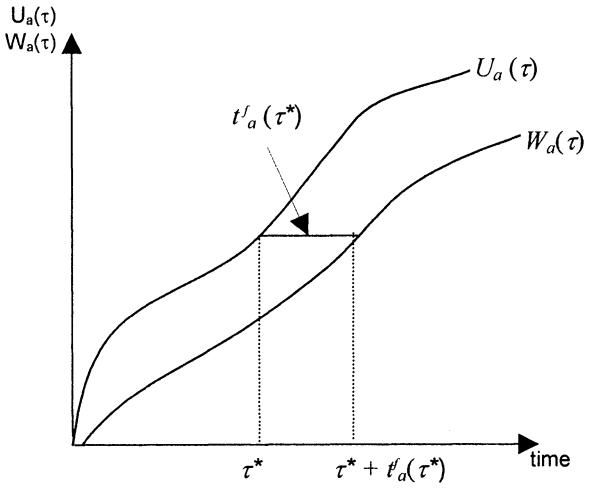


Fig. 6.2.8 Condition for FIFO rule (forward travel time).

Similar considerations can be derived analyzing equation (6.2.9b): the out-flow on a link at time τ is equal to the in-flow at the corresponding entrance time multiplied by a factor larger than one (i.e. the out-flow is bigger than the in-flow) if the flow on the link is accelerating ($dt/d\tau<0$) and vice versa.

Note that in the static case since link travel times are constant over time, in-flows always equal out-flows. Moreover, since a positive in-flow on a link, $u_a(\tau)>0$, implies a positive out-flow, $w_a(\tau)>0$, conditions (6.2.7) for FIFO discipline can be derived again from equations (6.2.9).

Once that the main variables and their consistency relationships have been introduced, the other components of the supply model, namely link performance functions, path performance functions and network flow propagation model can be analyzed.

6.2.1.2. Link performance and travel time functions

Fundamental to dynamic supply models are the link travel time functions expressing travel time as a function of link flows for congested networks. Most models proposed in the literature adopt functions simulating explicitly (i.e. *travel time functions*) or implicitly (i.e. *exit functions*) the travel time on a link depending on the number of users traveling on the link. Implicit exit time functions express directly the out-flow of a given link as a function of the link occupancy $w_a(\tau) = w_a(x_a(\tau))$. These functions, however, lead to a number of theoretical inconsistencies and will not be considered in the following.

Travel time functions express the travel time $t_a(\tau)$ of a particle arriving at the beginning of the link a at time τ as a function of the relevant traffic condition variables. Most models proposed in the literature adopt "separable" travel time functions, i.e. functions expressing the travel time $t^f_a(\tau)$ in terms of the instantaneous occupancy on the same link $x_a(\tau)$:

$$t_a^f(\tau) = t_a(x_a(\tau)) \tag{6.2.10}$$

The computation of the backward travel time function, or the inverse travel time function, from the forward function require the solution of a fixed-point problem whose solution is unique only if the FIFO rule applies. In fact, from equation (6.2.5b) it results:

$$t_a^b(\tau^*) = t_a^f(x_a(\tau - t_a^b(\tau^*))) \tag{6.2.11}$$

Several functional forms have been proposed for equation (6.2.10), not all of which however lead to results consistent with the FIFO rule. One of the proposed functions is the linear travel time function:

$$t_a^f(x_a(\tau)) = t_a^0 + \frac{1}{Q_a} \cdot x_a(\tau) \qquad (6.2.12a)$$

where Q_a is, as usual, the capacity of link a and t^0_a is the free-flow link travel time. It can be shown that a linear travel time function imposes the FIFO discipline and the consistency of the model. Furthermore the out-flow w_a never exceeds the capacity of link a. Fig. 6.2.9 plots out-flow as a function of the number of vehicles on the link.

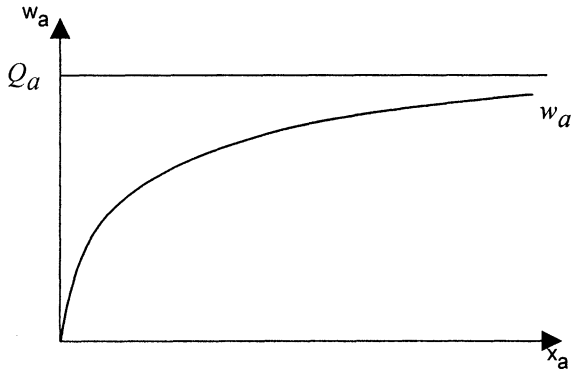


Fig. 6.2.9 Exit flow curve corresponding to linear travel time function (6.2.12a).

A similar function derived from deterministic queuing models can be applied for queuing links. In fact, all the concepts introduced so far apply to queuing models, see section 2.A.2. The only difference is that travel time is not spent moving on the link but waiting. In-flow and out-flow are the equivalent of arrival and departure flow rates, occupancy is equivalent to the number of queuing users and so on. In this case equation (6.2.12a) can be written as:

$$tw_a(x_a(\tau)) = \frac{1}{Q_a} + \frac{1}{Q_a} x_a(\tau) \qquad (6.2.12b)$$

where the "zero occupancy" time is equal to the average service time, i.e. $tw_a = 1/Q_a$.

6.2.1.3. Path performance and travel time functions

Specific time variables can also be associated with paths, (see Fig. 6.2.10). Let

$T^f_{a^k_i}(\tau)$ [or equivalently $T^f_{a^k}(\tau)$] be the forward travel time to link a_i [a] along path k; i.e. the time needed to reach the beginning of link a_i [a] following path k and leaving at time τ from the beginning of link a^k_1;

$T^b_{a^k_i}(\tau)$ [or equivalently $T^b_{a^k}(\tau)$] be the backward travel time necessary to reach the beginning of link a_i [a] following path k and arriving at time τ.

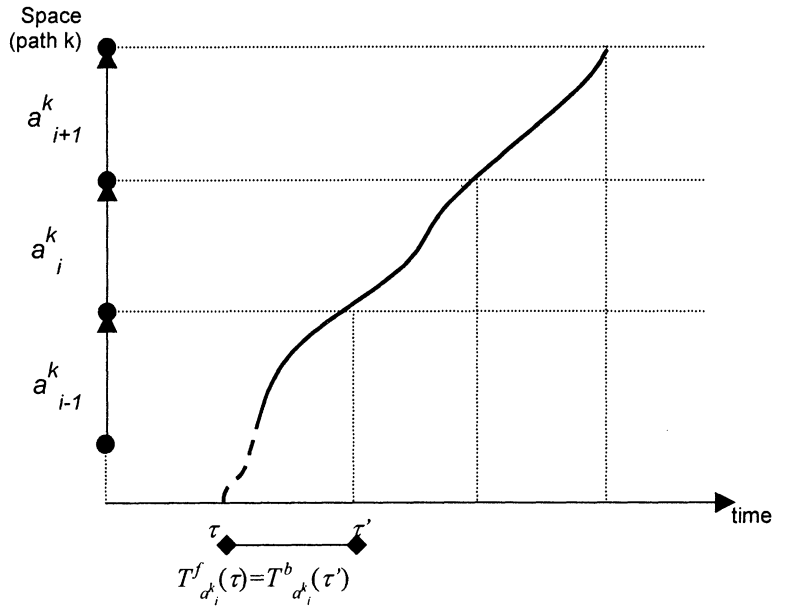


Fig. 6.2.10 Backward and forward path travel time.

Temporal consistency of the model requires that the different travel times satisfy the following relationships (see Fig. 6.2.10):

$$T^f_{a^k_i}(\tau) = T^b_{a^k_i}(\tau + T^f_{a^k_i}(\tau)) \tag{6.2.13a}$$

$$T^b_{a^k_i}(\tau) = T^f_{a^k_i}(\tau - T^b_{a^k_i}(\tau)) \tag{6.2.13b}$$

Moreover, let:

$TT^f_k(\tau)$ be the forward total travel time of path k, i.e. the time needed to traverse path k starting at time τ;

$TT^b_k(\tau)$ be the backward total travel time on path k, i.e. the time needed to traverse path k arriving at time τ;
$EC_k(\tau)$ be the path k generalized extra-cost starting at time τ;
$g_k(\tau)$ be the total generalized cost along path k leaving at time τ.

As for links, travel times along any path must satisfy FIFO conditions. This condition can be stated formally as (see Fig. 6.2.11):

$$\tau' + T^f_{a_i^k}(\tau') < \tau'' + T^f_{a_i^k}(\tau'') \qquad \forall \quad \tau' < \tau'' \tag{6.2.14a}$$

Similarly for the backward travel time:

$$\tau' - T^b_{a_i^k}(\tau') < \tau'' - T^b_{a_i^k}(\tau'') \qquad \forall \quad \tau' < \tau'' \tag{6.2.14b}$$

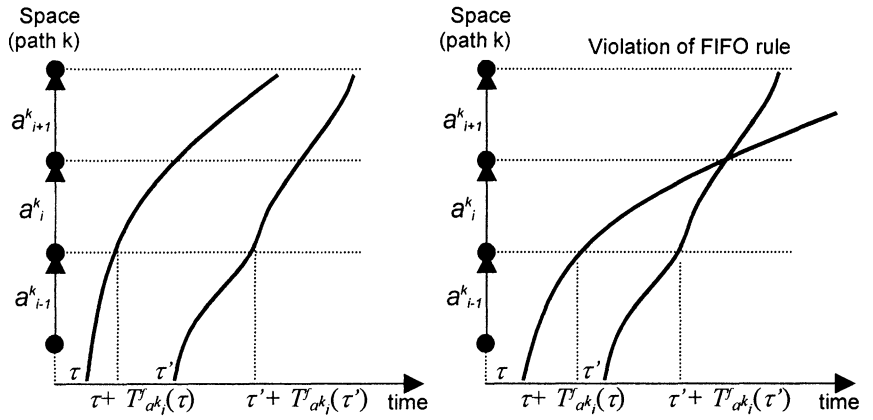


Fig. 6.2.11 Representation of the FIFO discipline on a path.

Relationships (6.2.14) imply that $\tau + T^f_{a_i^k}(\tau)$ and $\tau - T^b_{a_i^k}(\tau)$ are well-defined functions of τ, i.e. a single value corresponds to each value of τ. It can be shown easily that if all link travel time functions follow a FIFO discipline, the latter is also satisfied along a path.

Path and link travel times are connected through the following equations (see Fig. 6.2.12):

$$T^f_{a^k_{i+1}}(\tau) = T^f_{a_i^k}(\tau) + t^f_{a_i^k}(\tau + T^f_{a_i^k}(\tau)) \tag{6.2.15a}$$

$$T^b_{a^k_{i+1}}(\tau) = t^b_{a^k_i}(\tau) + T^b_{a^k_i}(\tau - t^b_{a^k_i}(\tau)) \qquad (6.2.15b)$$

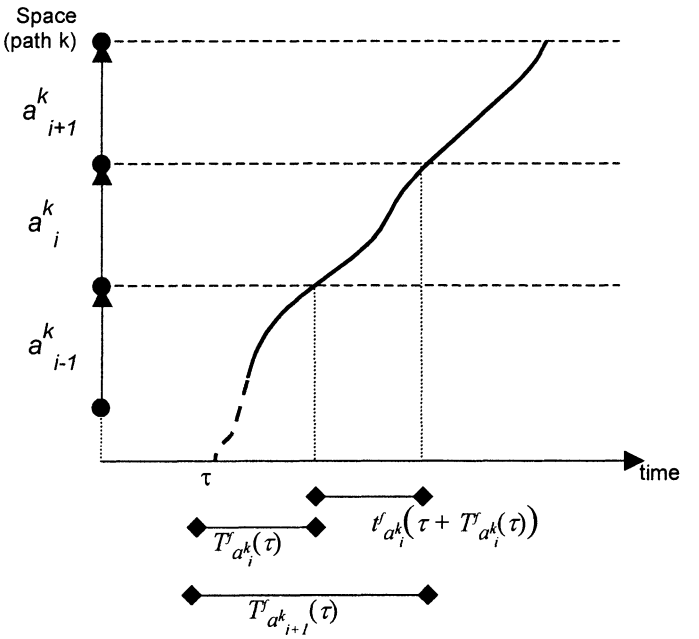


Fig. 6.2.12 Relationship between path and link travel times (forward travel time functions).

Equation (6.2.15a) can be applied recursively from the first link a^k_1 to the generic link a^k_i of path k. This results in a "nested" sum of link travel times:

$$T^f_{a^k_i}(\tau) = t^f_{a^k_1}(\tau) + t^f_{a^k_2}(\tau + t^f_{a^k_1}(\tau)) + t^f_{a^k_3}(\tau + t^f_{a^k_1}(\tau) + t^f_{a^k_2}(\tau + t^f_{a^k_1}(\tau))) + ...$$
$$+ t^f_{a^k_{i-1}}(\tau + t^f_{a^k_1}(\tau) + ... + t^f_{a^k_{i-2}}(\tau + t^f_{a^k_1}(\tau) + ...)) \qquad (6.2.16a)$$

Similarly equation (6.2.15b) can be applied from link a^k_i to the first link a^k_1 of path k:

$$T^b_{a^k_i}(\tau) = t^b_{a^k_{i-1}}(\tau) + t^b_{a^k_{i-2}}(\tau - t^b_{a^k_{i-1}}(\tau)) + t^b_{a^k_{i-3}}(\tau - t^b_{a^k_{i-1}}(\tau) - t^b_{a^k_{i-2}}(\tau - t^b_{a^k_{i-1}}(\tau))) + ...$$
$$... + t^b_{a^k_1}(\tau - t^b_{a^k_{i-1}}(\tau) - ... - t^b_{a^k_2}(\tau - t^b_{a^k_{i-1}}(\tau) - ... - t^b_{a^k_3}(\tau - t^b_{a^k_{i-1}}(\tau) - ...))) \qquad (6.2.16b)$$

Previous equations can be easily extended to express total path travel time, path extra-cost (assuming link-wise additive attributes) and generalized cost as functions of link travel times:

$$TT_k^f(\tau) = t_{a_1^k}^f(\tau) + t_{a_2^k}^f(\tau + t_{a_1^k}^f(\tau)) + \ldots + t_{a_{n_k}^k}^f(\tau + \ldots) = T_{a_{n_k}^k}^f(\tau) + t_{a_{n_k}^k}^f(\tau + \ldots) \quad (6.2.17a)$$

$$EC_k(\tau) = ec_{a_1^k}(\tau) + ec_{a_2^k}(\tau + t_{a_1^k}^f(\tau)) + \ldots + ec_{a_{n_k}^k}^f(\tau + \ldots) \quad (6.2.17b)$$

$$g_k(\tau) = \beta_t TT_k^f(\tau) + EC_k(\tau) \quad (6.2.17c)$$

The relationships between the vectors of forward path travel time functions $\boldsymbol{TT}^f(\tau)$, with one component for each path in the network and the vectors of forward link travel time functions $\boldsymbol{t}(\tau)$, with one component for each link in the network, can be expressed as:

$$\boldsymbol{TT}^f(\tau) = \Gamma(\boldsymbol{t}(\tau'), \tau' > \tau) \quad (6.2.18)$$

Equations (6.2.17) are the within-day dynamic equivalent of the link-wise cost composition expressed by supply model (2.2.5) and (5.2.1) for static networks. In the static case the order in which link performance attributes or costs are summed to obtain path costs is irrelevant. This is no longer true for within-day dynamic supply models in which link times and costs have to be summed up in their topological order along path k to satisfy the temporal succession of crossed links.

6.2.1.4. Dynamic Network Loading models

Dynamic Network Loading (DNL), also known as Dynamic Network Flow Propagation (DNFP), models simulate how time-varying continuous path flows propagate through the network inducing time-varying in-flows, out-flows and link occupancies.

The simplest case is that of a single-link network. The link flow propagation model can be expressed formally by combining the different consistency equations introduced in the previous section and the travel time function. In fact, the whole model expressing the continuous link flow dynamics is specified as a function of a single input variable, usually in-flow, since the four variables defining the dynamics of the link, namely $u_a(\tau)$, $w_a(\tau)$, $x_a(\tau)$ and $t_a^f(\tau)$, are connected by three equations:

$$\frac{dx_a(\tau)}{d\tau} = u_a(\tau) - w_a(\tau)$$

$$t_a^f(\tau) = t_a(x_a(\tau))$$

and, under the FIFO rule condition, such as:

$$u_a(\tau) = w_a(\tau + t_a^f(\tau)) \cdot \left(1 + \frac{dt_a^f(\tau)}{d\tau}\right)$$

The DNL model can be extended to general networks if the FIFO condition is satisfied by link and path travel times. In this case the conservation and link dynamics equations have to be stated with respect to specific path values:

$$u_{a_1^k}^k(\tau) = h_k(\tau) \qquad \forall k \qquad (6.2.19a)$$

$$u_{a_i^k}^k(\tau) = w_{a_{i-1}^k}^k(\tau) \qquad \forall k; \forall i = 1...n_{k-1} \qquad (6.2.19b)$$

$$\frac{dx_a^k(\tau)}{d\tau} = u_a^k(\tau) - w_a^k(\tau) \qquad \forall k; \forall a \qquad (6.2.19c)$$

$$x_a(\tau) = \sum_k \delta_{ak} x_a^k(\tau) \qquad \forall a \qquad (6.2.19d)$$

$$t_a^f(\tau) = t_a\left(x_a(\tau)\right) \qquad \forall a \qquad (6.2.19e)$$

$$w_a^k(\tau + t_a^f(\tau)) = \frac{u_a^k(\tau)}{\left(1 + \frac{dt_a^f(\tau)}{d\tau}\right)} \qquad \forall a \qquad (6.2.19f)$$

in addition to the boundary conditions (e.g. $u(0)=w(0)=x(0)=0$).

The above equations give an implicit representation of the DNL model; they can also be reformulated in such a way to bear a closer resemblance to their static counterpart, which can be shown to be a particular case. In fact, by applying equation (6.2.9b) to the whole path k, up to link a considered as a single link (see Fig. 6.2.13), it follows that:

$$u_a(\tau) = \sum_k \delta_{ak} \cdot h_k(\tau - T_{a^k}^b(\tau)) \cdot \left(1 - \frac{dT_{a^k}^b(\tau)}{d\tau}\right) \qquad (6.2.20)$$

Note that in the static case in-flows are constantly equal to out-flows, path flows are constantly equal to h_k and both link and path travel time are constant overtime:

$$u_a(\tau) = w_a(\tau) = u_a = w_a = f_a$$
$$h_k(\tau) = h_k$$
$$t_a^f(\tau) = t_a^b(\tau) = t_a$$
$$\frac{dt_a^f(\tau)}{d\tau} = \frac{dt_a^b(\tau)}{d\tau} = 0$$

Moreover, the fundamental diagram relations (see appendix 2.A) yields:

$$f_a = \frac{x_a}{L_a} \cdot v_a = \frac{x_a}{t_a}$$

Then the system of equations (6.2.19) becomes the linear system:

$$\boldsymbol{f}_a = \sum_k \delta_{ak} h_k \; ; \boldsymbol{f} = \Delta \cdot \boldsymbol{h}$$

expressing the network flow propagation model for static networks.

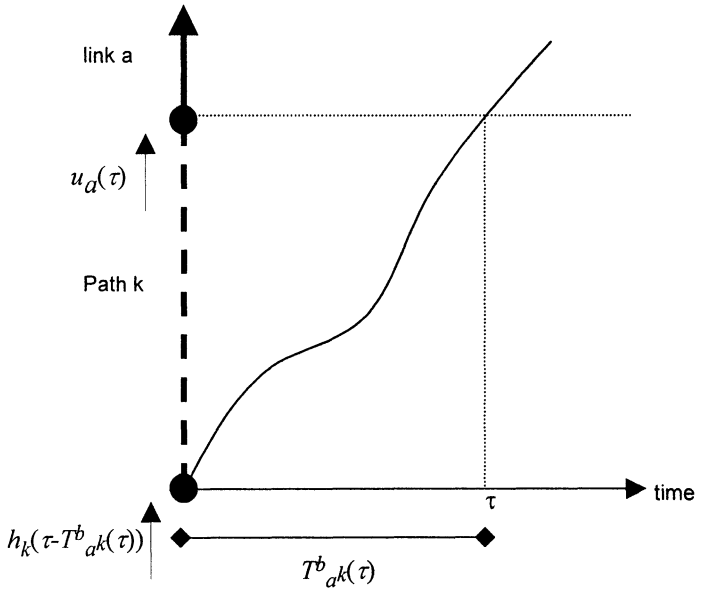


Fig. 6.2.13 Representation of the Dynamic Network Flow Propagation model.

On the other hand if path travel times are constant over time ($dT^b_{a,k} / d\tau = 0$), e.g. the network is uncongested and there are no changes in supply, the in-flow profile is the summation of the path-flow profiles for the paths including the link, shifted by the time needed to reach the link:

$$u_a(\tau) = \sum_k \delta_{ak} \cdot h_k(\tau - T^b_{a,k}(\tau)) \qquad (6.2.21)$$

6.2.1.5. Formalization of the overall supply model

The equations introduced in the previous sections express the dependence of in-flows, out-flows occupancies, link and path travel times and costs on path flows leaving in previous time instants. The relevant equations defining the overall supply model for congested networks respecting FIFO rule, can be expressed symbolically as:

$$\boldsymbol{f} = \Phi[\boldsymbol{t}(\tau), \boldsymbol{h}(\tau)] \tag{6.2.22a}$$

$$\boldsymbol{t}(\tau) = t(\boldsymbol{f}(\tau'), \tau' \leq \tau) \tag{6.2.22b}$$

$$\boldsymbol{TT}^{f}(\tau) = \Gamma(\boldsymbol{t}(\tau'), \tau' > \tau) \tag{6.2.22c}$$

where:

$\boldsymbol{t}(\tau)$ is the vector of link travel times at time τ;
$\boldsymbol{TT}(\tau)$ is the vector of forward path travel times at time τ;
$\boldsymbol{f}(\tau)$ denotes the vector of relevant flow or occupancy input variables for travel time functions at time τ;
$\boldsymbol{h}(\tau)$ is the path flow vector at time τ;
Γ expresses symbolically the relationship between link and path travel times, see equations (6.2.16);
Φ expresses symbolically the Dynamic Network Loading model, see equations (6.2.19);

These equations are the continuous flows within-day dynamic equivalent of the static equations:

$$\boldsymbol{f} = \Delta\, \boldsymbol{h}$$
$$\boldsymbol{c} = c(\boldsymbol{f})$$
$$\boldsymbol{g} = \Delta^{T}\, \boldsymbol{c}$$

Note that equations (6.2.22) reflect the fact that for congested networks the time to cross each link at a time τ depends on the flow on all the links of the network in the previous time τ', as it depends on the travel time to reach the link along the generic path k and, thus, depends on the travel time on links preceding a along each path k.

The solution of the dynamic supply model described is based on time discretization of the differential equation defining it. Given the large number of differential equations involved, the sequence in which they are processed is also relevant.

Note that this formulation of the supply model assumes that the relevant congestion variables influencing travel times are link-related occupancies at the time of arrival of a given particle at each link. This assumption, however convenient from

the computational point of view and "closer" to the static model, is appropriate only for deterministic queuing links and for very short running links.

Other continuous flow (i.e. *continuous space* models) are based on a direct application of differential equations systems derived from continuous space models of traffic flows for each link (see section 2.A.2), together with the equations assuring flow conservation at each node. The solution of these models, at least in theory, allows the definition of variables such as flow, speed, and density at each point s and at each instant τ. The solution of such models however requires a discretization in space, Δs, hence from the solution point-of-view they can be considered similar to discrete space model through a duly definition of link length (i.e. $\Delta s = L_a$).

6.2.2. Discrete flow supply models

Discrete flow models assume that users are discrete units; they can be either vehicles or groups of vehicles moving over the network and experiencing the same trip. In the following, discrete units will be referred to as *packets* including the special case of single-vehicle packets.

Discrete flow models require some form of discretization in time and can be based on two different approaches according to the way in which space is treated.

Mesoscopic models simulate the network performances at an aggregated level; as in the discrete-space continuous-flow models, aggregated variables of capacity, flows and occupancy are used. The traffic, however, is represented discretely by tracing the trips of single *packets*; each packet is characterized by a departure time and by a path up to the destination. It is usually assumed that the packets are concentrated at a point (concentrated or piled packets); this assumption is the more realistic, the smaller the size of the packets. Mesoscopic models can be applied to networks of general form and extended to simulate queue-formation and spill-backs with reasonable computing times. On the other hand, they do not allow a detailed simulation of traffic phenomena (overtaking, lane-changing, etc.).

Microscopic models explicitly simulate the time-space trajectory of each individual vehicle (speed, acceleration, etc.) and its interactions with nearby vehicles (overtaking, lane-changing, etc.) given the departure time from the origin and the path followed to the destination, as well as some individual characteristics (such as desired speed, driving style, etc.). These models (which in traffic theory literature are often referred to as *micro-simulation* models) can be very accurate. However, they do not allow the explicit formulation of the whole assignment model or the analysis of its theoretical properties. Moreover, it is very difficult to calibrate all the parameters of the model and considerable computational resources are required. Furthermore, this kind of models gives results at level of detail unnecessary for planning applications. Thus, the following will refer only to mesoscopic discrete flow models.

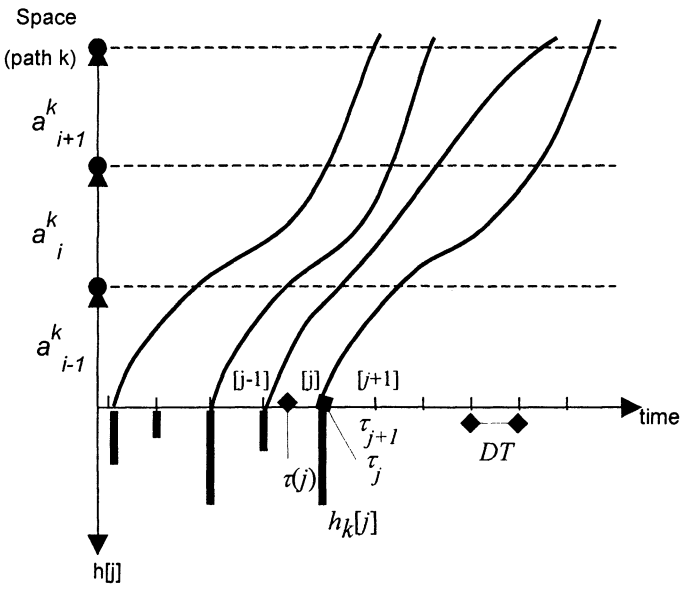


Fig. 6.2.14 Path flows and trajectories in discrete time-discrete flow models.

Most discrete flow models are based on some form of time discretization, i.e. divisions of the reference period into intervals $[j]$, (in the following intervals are assumed to be of equal duration DT for simplicity sake). These models often assume that relevant flow variables are averaged over time intervals. It is also assumed that users begin their trips at a characteristic time instant, τ_j, of an interval $[j]$. This may be the beginning or mid-point of the interval (see Fig. 6.2.14). In principle the duration of departure intervals can differ from the duration of averaging intervals. For example, some models use very short departure intervals while averaging the variables over longer intervals. In the following to simplify notation, only the single-interval case will be considered, the generalization to multiple intervals is rather straightforward. Furthermore it will be assumed that the representative instant of each interval is its final point, i.e. $\tau_j=[j]\cdot DT$.

A general framework for discrete time-discrete flow models is more difficult to formalize than for continuous models, since there are several possibilities to discretize the relevant variables. The framework proposed in the following is general enough to include several models proposed in the literature.

6.2.2.1. Variables and consistency conditions

Like continuous flow models, variables and their "structural " relationships must first be defined.

Time variables. The discretization of time requires the introduction of other time variables in addition to the generic absolute time τ. Let

$\tau(j)$ be the generic instant of time interval $[j]$, $\tau(j) \in ([j-1] \cdot DT, [j] \cdot DT)$;

τ_j be the characteristic instant of time interval $[j]$, here assumed to be its end-point, $\tau_j = [j] \cdot DT$.

Topological variables. Topological variables are the same as in the continuous-flow continuos-time case and will not be restated.

Flow and occupancy variables. The flow variables have the same definitions as in the continuous case, but in discrete flow models they represent "counts", i.e. number of users in a generic interval $[j]$, rather than flows, i.e. temporal densities, as shown in Fig. 6.2.14. In the following, however, they will be referred to indifferently as units (in a time interval) or flows to simplify the notation and the extension of continuous flow results. Let

k_j be the generic quantum or packet identified by the path k followed (and thus the O-D pair connected by k) and the departing interval $[j]$; only one packet can leave on a given path in each time interval;

$d_{od}[j]$ be the number of users moving between the pair od leaving in the representative instant of interval $[j]$;

$h_k[j]$ be the number of users starting their trip along path k, $k \in K_{od}$, in (the representative instant of interval $[j]$; $h_k[j]$ can be seen as the dimension of the packet k_j;

$f^k{}_{a,s}[j]$, $u^k{}_a[j]$, $w^k{}_a[j]$ be respectively the number of users moving on path k and crossing section s of link a, the number of users on path k entering and leaving link a during interval $[j]$.

$f_{a,s}[j]$, $u_a[j]$, $w_a[j]$ be respectively the total, summed over all path, number of users crossing section s of link a, the total number of users entering and leaving link a during interval $[j]$. Note that they correspond to the variables, introduced in section 2.A.1, $m(s|\tau_{j-1}, \tau_j)$ with symbols modified to parallel that used for continuous models.

Equations (6.2.1) expressing the total flows as sum of path flows and (6.2.2) expressing flow conservation at nodes hold also in the discrete case.

Flow variables can be defined also with respect to any sub-interval of interval j e.g. the interval $[\tau_{j-1}, \tau(j)]$ up to time $\tau(j)$ in this case they will be denoted as $f^k{}_{as}[\tau(j)]$ and so on. Let

$x_a(\tau_j)$, $x_a(\tau(j))$ be the link occupancy respectively in time instants τ_j and $\tau(j)$;

$\hat{x}_a[j]$ be the average occupancy on link a during interval $[j]$. It obviously results that:

$$\hat{x}_a[j] = \frac{1}{DT} \int_{[j-1]DT}^{[j]DT} x_a(\tau(j))d\tau(j)$$

$U_a[\tau_j]$ $U_a[\tau(j)]$, $W_a[\tau_j]$, $W_a[\tau(j)]$ be respectively the cumulated in-flows and out-flows of link a up to the representative instant of interval $[j]$ and to a generic time instant within that interval respectively. Cumulated in-flows and out-flows are related to interval specific values as:

$$U_a(\tau_j) = \sum_{j'<j} u_a[j] \tag{6.2.23a}$$

$$W_a(\tau_j) = \sum_{j'<j} w_a[j] \tag{6.2.23b}$$

In-flows and out-flows are also related to link occupancy through link conservation equations analogous to equations (6.2.3) and (6.2.4):

$$x_a(\tau_j) - x_a(\tau_{j-1}) = u_a[j] - w_a[j] \tag{6.2.24a}$$

$$x_a(\tau_j) = U_a(\tau_j) - W_a(\tau_j) \tag{6.2.24b}$$

Travel time and cost variables. In general, link and path travel times are continuous variables related to generic time instant τ as in the continuous case; in the discrete case, however, not all instants τ are meaningful since not all correspond to the arrival (or departure) of a packet (see Fig. 6.2.14). In the following time and cost variables will be introduced with respect to a generic instant τ. Let

$t^f_a(\tau)$, $t^b_a(\tau)$ be the forward and backward travel time on link a for a packet respectively entering or leaving the link at time τ. Forward and backward link travel times are related through mutual consistency equations identical to equations (6.2.5) which will be restated for the reader's convenience:

$$t^f_a(\tau) = t^b_a(\tau + t^f_a(\tau))$$

$$t^b_a(\tau) = t^f_a(\tau - t^b_a(\tau))$$

Since in discrete flow models users are identifiable units (packets k_j), it is possible to define temporal variables associated to the specific packet. Let:

$\tau^u_a[k_j]$, $\tau^w_a[k_j]$ be respectively the entrance and exit times on link a of packet k_j. Consistency with travel times requires that (see Fig. 6.2.15):

$$\tau^w_a[k_j] = \tau^u_a[k_j] + t^f_a(\tau^u_a[k_j]) \tag{6.2.25a}$$

$$\tau^u_a[k_j] = \tau^w_a[k_j] - t^b_a(\tau^w_a[k_j]) \tag{6.2.25b}$$

The FIFO discipline also applies to discrete models if it is assumed that packets cannot overtake each other, or if no explicit overtaking mechanism is introduced. The formal representation of the FIFO rule is identical to that for continuos flow models:

$$\tau' + t^f_a(\tau') < \tau'' + t^f_a(\tau'') \qquad \forall \quad \tau' < \tau''$$

and similarly for the backward travel time:

$$\tau' - t^b_a(\tau') < \tau'' - t^b_a(\tau'') \qquad \forall \quad \tau' < \tau''$$

Alternative conditions for the FIFO rule, analogous to those introduced in section 6.2.1.1 for continuous models can be stated. It should be observed, however, that, for discrete models, this condition is not so important since a packet is identified by the very nature of the model rather than implicitly through the trajectory crossing a given point at a given time.

As for the continuous case, the general discrete dynamic supply model can be formalized through link and path performance functions and the network flow propagation model.

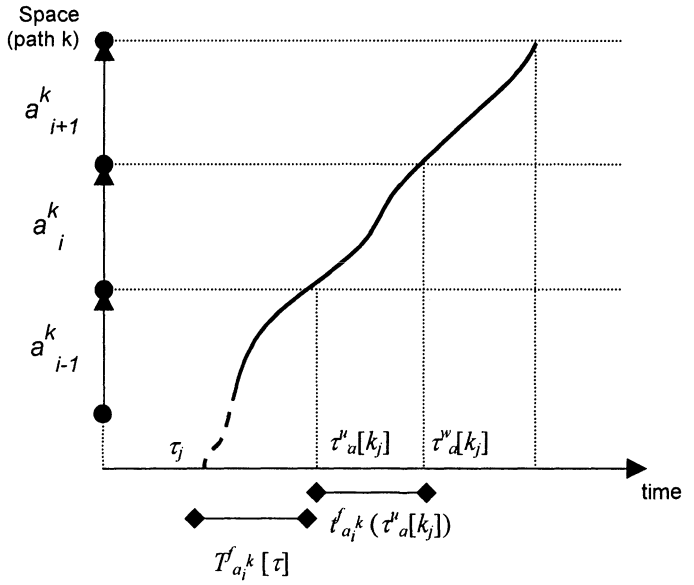


Fig. 6.2.15 Relationship between link entrance, exit and travel times of a packet on a link.

6.2.2.2. Link performance and travel time functions

The dependence of link travel time on link "flow" variables for congested networks can be expressed through a number of models. It is possible to specify separable and non-separable cost functions, the latter possibly allowing for spill-back effects from downstream links. The simpler separable travel time functions are similar to the functions adopted for running and queuing links described in section 6.2.1.2.

Forward travel time on running link a can be expressed as a linear function of arrival time, thus varying for different time instants $\tau(j)$ within interval j:

$$t_a^f(\tau(j)) = t_a^0 + \frac{1}{Q_a} \cdot x_a(\tau(j)) \tag{6.2.26}$$

Other models express the travel time via the average speed computed as a function of link density as in the fundamental diagram of traffic flow described in Section 2.A:

$$t_a^f(\tau(j)) = \frac{L_a}{V_a(x_a(\tau(j))/L_a)} \tag{6.2.27}$$

Given the discrete nature of the models, several assumptions can be made on the computation of travel times for packets entering the link in a given interval.

Some models proposed in the literature assume that the travel times are equal for all packets entering the link in a given interval. In this case occupancy variable in equations (6.2.26) and (6.2.27) correspond to a representative time of interval j, typically its start-point, τ_{j-1} and are constant for all users entering the link during the interval. Alternatively travel times can be computed as functions of the average link occupancy during the previous interval, $\hat{x}_a[j-1]$ or the same interval, $\hat{x}_a[j]$; in the latter case, however, link travel time for users entering the link during the interval depends on users entering the link later in the same interval. This may cause inconsistencies and counter-intuitive results and should be avoided. Other more accurate models compute travel times for each packet, e.g. as a function of the instantaneous link occupancy at the entrance time.

6.2.2.3. Path performance and travel time functions

The concepts of foreword and backward travel time needed to reach link a^k_i along path k leaving or arriving in a given instant can be immediately extended to discrete supply models. These variables will be denoted by $T^f_{a_i^k}(\tau_j)$ and $T^b_{a_i^k}(\tau(j))$ respectively to stress the fact that departures can occur only at the representative time of each interval, τ_j, while arrivals can be at any time during the interval $\tau(j)$, see Fig. 6.2.15. Equations (6.2.13) expressing the relationships between forward and backward travel times apply also to the discrete case. Similarly the forward (backward) total travel time on path k for a given departure (arrival) time can be defined also for the discrete case, denoting the variables with $TT^f_k(\tau_j)$and $TT^b_k(\tau(j))$ respectively.

The FIFO rule for partial and total path travel times can also be extended to discrete flow models, see equations (6.2.14).

Similarly the relationship between link and path travel times is analogous to equation (6.2.15) and, when applied recursively, leads to a "nested" structure corresponding to equation (6.2.16):

$$T^f_{a_i^k}(\tau_j) = t^f_{a_1^k}(\tau_j) + t^f_{a_2^k}(\tau_j + t^f_{a_1^k}(\tau_j)) + t^f_{a_3^k}(\tau_j + t^f_{a_1^k}(\tau_j) + t^f_{a_2^k}(\tau_j + t^f_{a_1^k}(\tau_j))) + \ldots$$
$$+ t^f_{a_{i-1}^k}(\tau_j + t^f_{a_1^k}(\tau_j) + \ldots\ldots + t^f_{a_{i-2}^k}(\tau_j + t^f_{a_1^k}(\tau_j) + \ldots\ldots)) \qquad (6.2.28)$$

In the discrete flow case, however, equation (6.2.28) can be expressed more straightforwardly by using the arrival times of the generic packet to link a^k_i, $\tau^u_{a_i^k}[k_j]$, as:

$$T^f_{a_i^k}(\tau_j) = t^f_{a_1^k}(\tau_j) + t^f_{a_2^k}(\tau^u_{a_2^k}[k_j]) + t^f_{a_3^k}(\tau^u_{a_3^k}[k_j]) + \ldots\ldots\ldots + t^f_{a_{i-1}^k}(\tau^u_{a_{i-1}^k}[k_j]) \qquad (6.2.29)$$

The same construct applies to total path travel time $TT^f_k(\tau_j)$, and to other path-additive attributes $EC_k(\tau_j)$ and finally to total path cost $g_k(\tau_j)$:

$$TT_k^f(\tau_j) = t^f_{a_1^k}(\tau_j) + t^f_{a_2^k}(\tau^u_{a_2^k}(\tau_j[k_j]) + t^f_{a_3^k}(\tau^u_{a_3^k}[k_j]) + \ldots$$
$$\ldots + t^f_{a_{n_k}^k}(\tau^u_{a_{n_k}^k}[k_j]) = T^f_{a_{n_k}^k}(\tau_j) + t^f_{a_{n_k}^k}(\tau^u_{a_{n_k}^k}[k_j]) \quad (6.2.30a)$$

$$EC_k(\tau_j) = ec^f_{a_1^k}(\tau_j) + ec^f_{a_2^k}(\tau^u_{a_2^k}[k_j]) + ec^f_{a_3^k}(\tau^u_{a_3^k}[k_j]) + \ldots$$
$$\ldots + ec^f_{a_{n_k}^k}(\tau^u_{a_{n_k}^k}[k_j]) \quad (6.2.30b)$$

$$g_k(\tau_j) = \beta_t TT_k^f(\tau_j) + EC_k(\tau_j) \quad (6.2.30c)$$

Formally the relationship between the vector of total path travel time, $TT^f(\tau_j)$, for a given departure time τ_j and travel times on the links making up each path, can be expressed symbolically as:

$$\boldsymbol{TT}^f(\tau_j) = \Gamma(\boldsymbol{t}(\tau'), \tau' \geq \tau_j) \quad (6.2.31)$$

Equation (6.2.31) is the equivalent of equation (6.2.18) in the continuos-flow case.

6.2.2.4. Dynamic Network Loading models

Unlike the continuous-flow case, the DNL model for discrete flows can be formulated explicitly since packets can be identified while moving over the network. In this case the in-flow on link a in the interval $[j]$ can be expressed as:

$$u_a[j] = \sum_k \sum_{l \leq j} \delta_{ak}[l, j] \cdot h_k[l] \quad (6.2.32)$$

where the $\delta_{ak}(l,j)$ are zero/one variables analogous to the elements of the static link-path incidence matrix; they are equal to one if the packet k_l (of intensity h_k $[l]$) enters link a during interval j, 0 otherwise:

$$\delta_{ak}[l, j] = \begin{cases} 1 & if \quad \tau^u_a[k_l] \in ([j-1]DT, [j]DT) \\ 0 & otherwise \end{cases}$$

Obviously the $\delta_{ak}[l,j]$ are all equal to zero if link a does not belong to path k (compare equation 6.2.30 with equations 6.2.20 and 6.2.21).

Equation (6.2.32) can also be formulated using a matrix notation as:

$$\boldsymbol{u}[j] = \sum_{l \le j} \Delta[l, j] \cdot \boldsymbol{h}[l] \tag{6.2.33}$$

which is close to the static counterpart $\boldsymbol{f} = \Delta \boldsymbol{h}$.

Similar equations can be stated for the out-flow, $w_a[j]$, from the generic link a at time interval j:

$$w_a[j] = \sum_k \sum_{l \le j} \delta'_{ak}[l, j] \cdot h_k[l] \tag{6.2.34}$$

where the $\delta'_{ak}[l,j]$ is equal to one if the packet k_l (of intensity h_k $[l]$) leaves link a during interval j, 0 otherwise:

$$\delta'_{ak}[l, j] = \begin{cases} 1 & if \quad \tau_a^w[k_l] \in ([j-1]DT, [j]DT) \\ 0 & otherwise \end{cases}$$

and in matrix terms:

$$\boldsymbol{w}[j] = \sum_{l \le j} \Delta'[l, j] \cdot \boldsymbol{h}[l] \tag{6.2.35}$$

Note that the elements of dynamic incidence matrices depend on link travel times and, for congested networks, on link flows and occupancies. In this respect they should be denoted as:

$$\delta_{ak}[l, j] = \delta_{ak}[l, j]\left(\boldsymbol{t}(\tau'); \tau' \in (\tau_l, \tau_j)\right)$$

The overall DNL model relating link flows and occupancies to path flows can be expressed combining the previous equations:

$$x_a(\tau_j) - x_a(\tau_{j-1}) = u_a[j] - w_a[j] \tag{6.2.36a}$$

$$u_a[j] = \sum_{l \le j} \Delta[l, j] \cdot \boldsymbol{h}[l] \tag{6.2.36b}$$

$$w_a[j] = \sum_{l \le j} \Delta'[l, j] \cdot \boldsymbol{h}[l] \tag{6.2.36c}$$

$$\tau_{a^k}^u[k_l] = \tau_l + T_{a_i^k}^f(\tau_l) \tag{6.2.36d}$$

$$\tau_{a^k}^w[k_l] = \tau_{a^k}^u[k_l] + t_{a^k}^f(\tau_{a^k}^u[k_l]) \tag{6.2.36e}$$

$$t_a^f(\tau(j)) = \frac{L_a}{V_a\left(x_a(\tau_{j-1})/L_a\right)} \tag{6.2.36f}$$

The above set of equations has been specified under the assumption that link travel time functions depend on link occupancy at the beginning of each interval; the model can be expressed in a similar form with reference to a generic time instant $\tau(j)$.

6.2.2.5. Formalization of the overall supply model

Equations (6.2.36) can be expressed symbolically as non-linear vector functions relating link flows (in-flows and out-flows) and occupancies for an interval j, to the vector of path flows leaving in intervals from l to j and the link travel times successive to τ_l and previous to the end of interval j, τ_j:

$$\boldsymbol{f}[j] = \Phi\left(\boldsymbol{h}[l], \boldsymbol{t}(\tau'); l \le j, \tau' \in [\tau_l, \tau_j]\right) \tag{6.2.37a}$$

Expression (6.2.37a) can be further combined with the equation relating link travel times to link occupancies for congested dynamic network loading models:

$$\boldsymbol{f}[j] = \Phi\left(\boldsymbol{h}[l], \boldsymbol{t}(\boldsymbol{x}(\tau')); l \le j, \tau' \in [\tau_l, \tau_j]\right) \tag{6.2.37b}$$

The global supply model is completed by the symbolic relationships relating path travel times to link travel times:

$$\boldsymbol{TT}^f(\tau_l) = \Gamma(\boldsymbol{t}(\boldsymbol{x}(\tau')); \tau' \ge \tau_l) \tag{6.2.38}$$

and path generalized transportation costs to travel times and other link costs vectors:

$$\boldsymbol{g}(\tau_l) = \beta_t \boldsymbol{TT}^f(\tau_l) + \boldsymbol{EC}(\tau_l) \tag{6.2.39}$$

6.3. Demand models for continuous service systems

Demand models used in dynamic assignment express the relationship between path flows and path costs. The "minimal" demand model, i.e. the model included in all assignment models, relates to path and departure time choice; it will be described in this section. Other models simulating users learning and choice adjustment mechanisms needed for dynamic process assignment will be briefly described in the next section on demand-supply interaction.

The flow of users following a path k connecting the O-D pair od and starting at time τ, $h_k(\tau)$ can be represented with *elastic demand profile* models, simulating in addition to path, departure time choice given the desired arrival time at destination, τ_d, or the desired departure time from the origin τ_o.

The continuous time-continuous flow model will be discussed first. Let:

$d_{od}(\tau_d)$ be the flow of trips between the pair od with desired arrival time τ_d;
$p_{od,k}(\tau/\tau_d)$ be the choice probability of time τ and path k, given the O-D pair od and the desired arrival time τ_d;
$V_k(\tau/\tau_d)$ be the systematic utility of path k and departure time τ, given the desired arrival time τ_d;
$V_{od}(\tau/\tau_d)$ be the vector of systematic utilities relative to all the paths connecting the pair od, $k \in K_{od}$, for a given departure time τ and desired arrival time τ_d.

The demand conservation condition over the whole reference interval $[O,T]$ can be formally expressed as:

$$h_k(\tau) = \int_0^T d_{od}(\tau_d)\, p_{od,k}(\tau / \tau_d)\, d\tau_d \tag{6.2.40}$$

Choice probabilities of departure time τ and path k are usually expressed with random utility models as a function of the systematic utilities of available path-departure time alternatives:

$$p_{od,k}(\tau/\tau_d) = p_{od,k}(V_{od}(\tau'/\tau_d),\ \forall\tau') \tag{6.2.41a}$$

Such models are usually "single-level" random utility models (e.g. Multinomial Logit) with mixed continuous (departure time)/discrete (path) alternatives, or partial share models. A sequence, which is sometimes used, is the product of path choice given the departure time, and the departure time choice models:

$$p_{od,k}(\tau/\tau_d) = p_{od}(\tau/\tau_d)\, p_{od}[k/\tau, \tau_d] \tag{6.2.41b}$$

Some empirical results related to demand elasticities with respect to changes in departure time and path seem to suggest a different sequence:

$$p_{od,k}(\tau / \tau_d) = p_{od}[k] \cdot p_{od}(\tau / k, \tau_d) \qquad (6.2.41c)$$

The specification of a simultaneous Multinomial Logit model for equation (6.2.41a) is:

$$p_{od,k}(\tau / \tau_d) = \frac{exp(V_k(\tau / \tau_d))}{\sum_{k' \in K_{od}} \int_0^T exp(V_{k'}(\tau' / \tau_d))d\tau'}$$

Some dynamic assignment models proposed in the literature assume deterministic utility departure time and path models. In this case, as for static systems, choice probabilities cannot be expressed in closed form as may exist several departure time/path alternatives with equal systematic disutilities. Indirect expressions similar to the static models described in Chapter 4 can be adopted:

$$p_{od,k}(\tau/\tau_d)>0 \Rightarrow V_{od,k}(\tau/\tau_d) \geq V_{od,k}(\tau'/\tau_d) \ \forall \tau', k'$$

Deterministic choice models, however, are less realistic than they are in the static case when applied to continuous departure times.

Systematic utility functions proposed for the simulation of the combined path-departure time choice include, in addition to path attributes, the *schedule delay*, i.e. the penalty for arriving early or late with respect to the desired arrival time (see Fig. 6.2.19). In case of desired arrival time τ_d, it results:

$$V_k(\tau/\tau_d)=\beta_t TT^f_k(\tau) + EC_k(\tau)+\beta_e EAP_k(\tau, \tau_d, TT_k(\tau))+\beta_l LAP_k(\tau, \tau_d, TT_k(\tau)) \quad (6.2.42a)$$

where:

$EAP_k(\tau, \tau_d, TT^f_k(\tau))$ is the penalty related to early arrival with respect to τ_d departing in τ and following path k. This penalty is usually considered only if the early arrival is above a minimum threshold Δ_e:

$$EAP_k(\tau, \tau_d, TT^f_k(\tau)) = \tau_d - \Delta_e - (\tau + TT^f_k(\tau)) \quad \text{if} \quad \tau_d - \Delta_e - (\tau + TT^f_k(\tau)) > 0$$

$$= 0 \qquad \textit{otherwise}$$

$LAP_k(\tau, \tau_d, TT^f_k(\tau))$ is the penalty related to a delay with respect to τ_d departing in τ and following path k. This penalty is usually considered only if the delay is above a minimum threshold Δ_l:

$$LAP_k [\tau, \tau_d, TT^f_k(\tau)] = \tau + TT^f_k(\tau) - \tau_d - \Delta_l \qquad \text{if} \quad \tau + TT^f_k(\tau) - \tau_d - \Delta_l > 0$$

$$= 0 \qquad \textit{otherwise}$$

In case of desired departure time from the origin τ_o, the expression of the systematic utility is still a function of path travel time and scheduled delay, but in this case the scheduled delay does not depend on the path travel time $TT^f_k(\tau)$:

$$V_k(\tau/\tau_o) = \beta_t TT^f_k(\tau) + EC_k(\tau) + \beta_e EDP(\tau, \tau_o) + \beta_l LDP(\tau, \tau_o) \qquad (6.2.42b)$$

where:

$EDP(\tau, \tau_o)$ is the penalty related to early departure with respect to τ_o departing in τ, usually considered only if the early departure is above a minimum threshold Δ_e:

$$EDP(\tau, \tau_o) = \tau_o - \Delta_e - \tau \qquad \text{if } \tau_o - \Delta_e - \tau > 0$$

$$= 0 \qquad \textit{otherwise}$$

$LDP(\tau, \tau_o)$ is the penalty related to a delay with respect to τ_o departing in τ, usually considered only if the delay is above a minimum threshold Δ_l:

$$LDP(\tau, \tau_o) = \tau - \tau_o - \Delta_l \qquad \text{if } \tau - \tau_o - \Delta_l > 0$$

$$= 0 \qquad \textit{otherwise}$$

All the coefficients β in equations (6.2.42) are negative. Furthermore, the schedule delay penalties should have coefficients β_e and β_l with absolute values greater than the travel time coefficient ($|\beta_e| > |\beta_t|$, $|\beta_l| > |\beta_t|$) in order to avoid unrealistic user behavior, e.g. large probabilities for alternatives with very high early/late arrival penalties but with smaller travel times. Empirical results for work related trips show that the disutility of late arrivals is larger than that for early arrivals ($|\beta_e| < |\beta_l|$), as shown in Fig. 6.3.1.

The global within-day dynamic demand model with elastic demand profile is expressed by equations (6.2.40), (6.2.41) and (6.2.42) relating path flows to path travel times, extra costs and schedule delay penalties for different departure times.

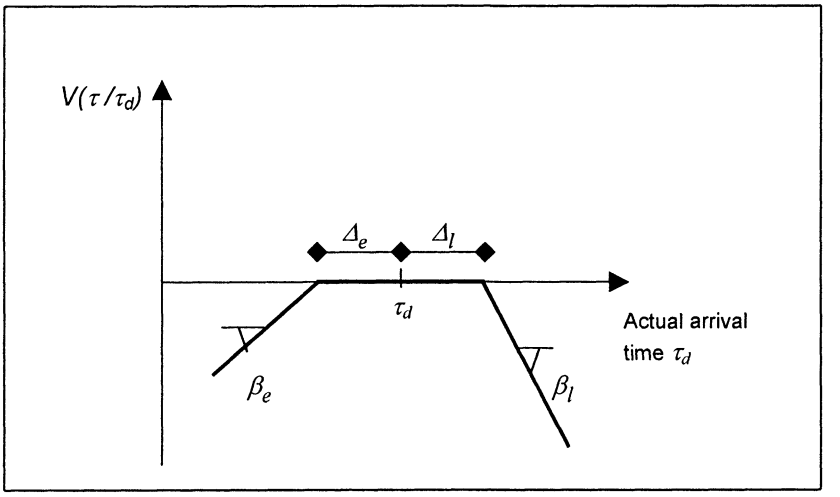


Fig. 6.3.1 Systematic utility function with respect to desired arrival time.

In *rigid demand profile* models, it is assumed that the distribution of demand flows over departure times is known and independent from variations in travel times, i.e. the probabilities $p_{od}(\tau/\tau_d)$ or $p_{od}(\tau/\tau_o)$ are given. It follows that path is the only choice dimension considered given a departure time:

$$h_k(\tau) = d_{od}(\tau)\, p_{od,k}(\boldsymbol{V}_{od}(\tau)) \tag{6.2.43}$$

where:

$d_{od}(\tau)$ is the demand flow leaving in at time τ;
$p_{od,k}(\tau)$ is the probability of choosing path k for trips starting at time τ;
$\boldsymbol{V}_{od}(\tau)$ is the vector of the systematic utilities $V_k[\tau]$ of the different paths connecting the O-D pair od, $k \in K_{od}$.

In this case path choice models are analogous to those described in Section 4.3.4; the systematic utility of the path k can be expressed as a function of the path-related attributes introduced previously as:

$$V_k(\tau) = \beta_t\, TT^f_k(\tau) + EC_k(\tau) \tag{6.2.44}$$

The within-day dynamic demand model with a rigid demand profile is expressed by the equations (6.2.43) and (6.2.44) connecting path flows to path travel times for a given departure time τ.

The extension of dynamic demand models to the discrete case is rather straightforward. The only difference is that alternative departure times are the discrete intervals ..$[j-1]$, $[j]$, $[j+1]$, or representative instants ... τ_{j-1}, τ_j, τ_{j+1}... . Simultaneous departure time and path choice probabilities are thus expressed as $p_{od,k}[\tau_j/\tau_d]$. A multinomial Logit specification can be:

$$p_{od,k}[\tau_j / \tau_d] = \frac{exp\left(V_k[\tau_j / \tau_d]\right)}{\sum_{\tau_{j'}} \sum_{k' \in K_{od}} exp\left(V_{k'}[\tau_{j'} / \tau_d]\right)}$$

Alternatively a partial share specification similar to equation (6.2.41b) introducing a correlation structure among adjacent departure intervals, e.g. with a Cross-Nested Logit model, can be adopted.

The previous results for choice models and systematic utility specifications apply also to the discrete departure time case. Discrete departure time models can be adopted for the continuous flows. In fact, some specifications of continuous departure time choice model assume that travelers do not choose among an infinite number of departure instant, but rather among a finite number of times intervals (e.g. 5 minutes long), and that actual departure times are uniformly distributed within the chosen interval. In this case the probability of leaving at time $\tau(j)$, within interval j, and following path k computed with a Multinomial Logit model would be:

$$p_{od,k}(\tau(j)/\tau_d) = \frac{1}{DT} \frac{\exp\left(V_k[j / \tau_d]\right)}{\sum_{j'} \sum_{k' \in K_{od}} \exp\left(V_{k'}[j' / \tau_d]\right)}$$

6.4. Demand-supply interaction models for continuous service systems

Demand-supply interaction models for within-day dynamic continuous services systems are conceptually analogous to those described for the equivalent static systems.

In the following sections some formal results will be given for uncongested network assignment as well as for congested network which can be approached through equilibrium or dynamic process models. The various models will be described making reference to the continuos and discrete flow cases, at least for uncongested and users equilibrium assignment models; on the other hand dynamic process models, with and without information, will be formulated only for discrete flow models.

Dynamic Traffic Assignment (DTA) models are rather complex and few operational formulations have been developed. Furthermore, there are no general theoretical results to date for the analysis of existence and uniqueness of the resulting flow configurations analogous to those obtained for static models.

For simplicity, in the following only (within-day dynamic) demand models with desired departure time τ_o will be illustrated. The extension to the case of desired arrival time is rather straightforward.

6.4.1. Uncongested Network assignment models

Dynamic assignment models for uncongested networks can be represented schematically as in Fig. 6.4.1. In this case link travel times do not depend on link occupancies.

In the continuous-flow case, they can be specified as:

$$\boldsymbol{t}^f(\tau) = \boldsymbol{t}^0(\tau) \tag{6.4.1a}$$

$$\boldsymbol{TT}^f(\tau) = \Gamma(\boldsymbol{t}^0(\tau)) \tag{6.4.1b}$$

$$V_{od}(\tau / \tau_o) = \beta_t \boldsymbol{TT}(\tau) + \boldsymbol{EC}(\tau) + \beta_e \boldsymbol{EDP}(\tau, \tau_o) + \beta_l \boldsymbol{LDP}(\tau, \tau_o) \tag{6.4.1c}$$

$$\boldsymbol{h}(\tau) = \sum_{\tau_o} \boldsymbol{P}\left(V_{od}(\tau/\tau_o)\right) \cdot \boldsymbol{d}(\tau_o) \tag{6.4.1d}$$

$$\boldsymbol{f}(\tau) = \Phi\left(\boldsymbol{h}(\tau), \boldsymbol{t}^0(\tau)\right) \tag{6.4.1e}$$

Equations (6.4.1c) and (6.4.1d) represent the within-day dynamic demand models. On the other hand, equations (6.4.1a) (6.4.1b) and (6.4.1e) make up the supply model representing respectively the link performance model, the path

performance model and the dynamic network flow propagation model. The uncongested dynamic assignment model (UND) can be deterministic (DUND) or stochastic (SUND) depending on the path choice model used in equation (6.4.1.d).

The Dynamic Network Loading model (DNL) has been formulated symbolically in terms of a characteristic link flow vector, $\boldsymbol{f}$, since, if FIFO rule holds, the different formulations in terms of in-flow, out-flow or link occupancy are equivalent. For instance, equation (6.4.1.e) can be stated in terms of in-flows as (see 6.2.1.4):

$$u_a(\tau) = \sum_k \delta_{ak} \cdot h_k(\tau - T^b_{a,k}(\tau))$$

where, since the network is uncongested, the backward travel times $T^b_{a,k}$ are independent on flows for each link a, but in general depend on the specific time τ:

$$T^b_{a,k}(\tau) = T^0_{a,k}(t^0(\tau)) \qquad (6.4.1f)$$

From equations (6.4.1) it results that in principle both demand and link travel times vary with τ. However, given the absence of congestion, equations (6.4.1) can be solved sequentially to obtain path performances and link flows. In uncongested networks it is usually assumed that link travel times are constant with respect to τ, i.e. $t^f_a(\tau)=t^0_a$. Thus the system of equations (6.4.1) becomes:

$$\boldsymbol{t}^f(\tau) = \boldsymbol{t}^0 \qquad (6.4.2a)$$

$$\boldsymbol{TT}^f(\tau) = \Gamma(\boldsymbol{t}^0) \qquad (6.4.2b)$$

$$\boldsymbol{V}_{od}(\tau/\tau_o) = \beta_t \boldsymbol{TT}(\tau) + \boldsymbol{EC}(\tau) + \beta_e \boldsymbol{EDP}(\tau,\tau_o) + \beta_l \boldsymbol{LDP}(\tau,\tau_o) \qquad (6.4.2c)$$

$$\boldsymbol{h}(\tau) = \sum_{\tau_o} \boldsymbol{P}\left(\boldsymbol{V}_{od}(\tau/\tau_o)\right) \cdot \boldsymbol{d}(\tau_o) \qquad 6.4.2d)$$

$$\boldsymbol{f}(\tau) = \Phi\left(\boldsymbol{h}(\tau), \boldsymbol{t}^0\right) \qquad (6.4.2e)$$

Here the only elements varying within-day are the demand flows inducing time-varying path and link flows. In particular equation (6.4.2b) becomes:

$$TT_k(\tau) = \sum_k \delta_{ak} \cdot t^0_a \qquad \forall \tau$$

or

$$\boldsymbol{TT}(\tau) = \Delta^T \cdot \boldsymbol{t}^0 \qquad \forall \tau$$

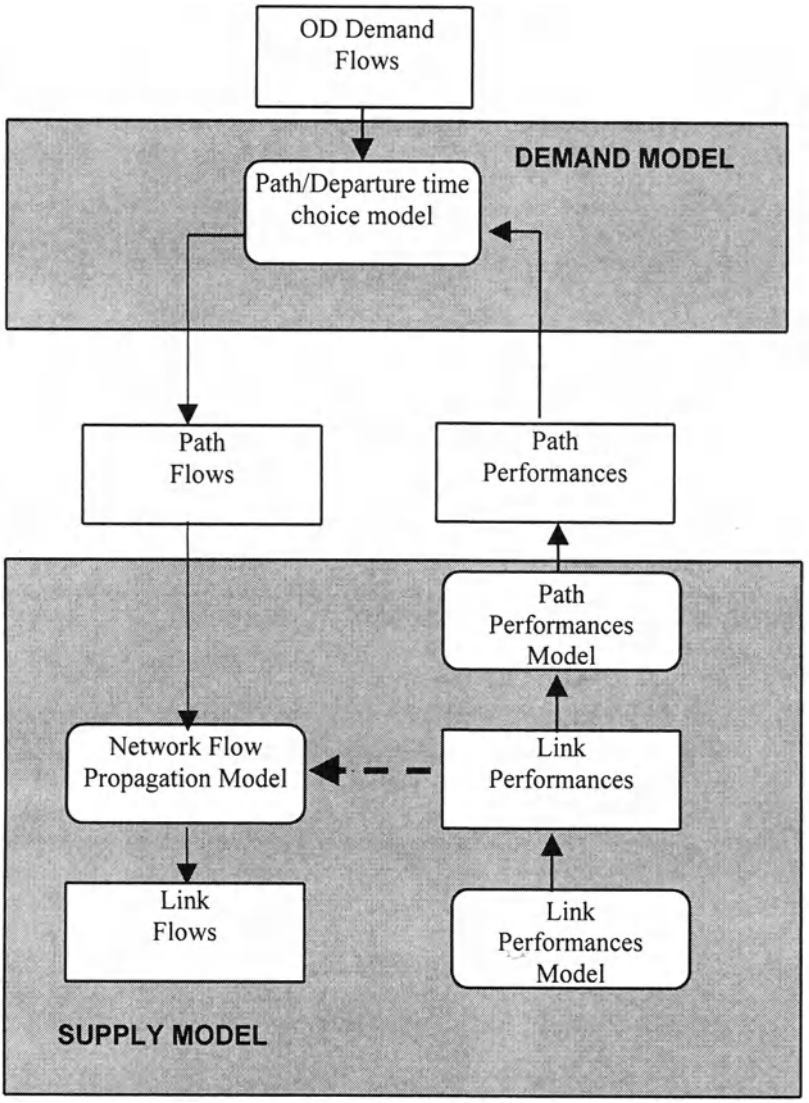


Fig. 6.4.1 Within-day Dynamic Traffic Assignment for uncongested networks.

In the discrete-flow case, the uncongested network assignment models can be formally specified as:

$$\boldsymbol{t}^{f}(\tau_j) = \boldsymbol{t}^{0}_{j} \tag{6.4.3a}$$

$$\boldsymbol{TT}^{f}(\tau_j) = \Gamma(\boldsymbol{t}^{0}_{j}) \tag{6.4.3b}$$

$$\boldsymbol{V}_{od}(\tau_j/\tau_o) = \beta_t \boldsymbol{TT}^{f}(\tau_j) + \boldsymbol{EC}(\tau_j) + \beta_e \boldsymbol{EDP}(\tau_j, \tau_o) + \beta_l \boldsymbol{LDP}(\tau_j, \tau_o) \tag{6.4.3c}$$

$$\boldsymbol{h}(\tau_j) = \sum_{\tau_o} \boldsymbol{P}\left(\boldsymbol{V}_{od}(\tau_j/\tau_o)\right) \cdot \boldsymbol{d}(\tau_o) \tag{6.4.3d}$$

$$\boldsymbol{f}[j] = \Phi\left(\boldsymbol{h}(\tau_j), \boldsymbol{t}^{0}_{j'}; j' < j\right) \tag{6.4.3e}$$

Note that in the above equations time dependency can be expressed equivalently as the representative time instant of interval j, τ_j, or simply as $[j]$. Note also that in the assignment model the packets introduced in the DNL model described in section 6.2.2.4 are the result of the departure-time/path choice.

Equations (6.4.3c) and (6.4.3d) represent the within-day dynamic demand models while equations (6.4.3a) (6.4.3b) and (6.4.3e) represent respectively the link performance model, the path performance model and the dynamic network flow propagation model, components of the overall supply model. The DNL can also be stated as:

$$\boldsymbol{f}[j] = \sum_{l \le j} \Delta[l, j] \cdot \boldsymbol{h}[l]$$

Note that, if link travel times are constant for all time intervals of the simulation period, i.e. $t^{f}_{a}(\tau) = t^{0}_{a}$, the matrix Δ does not depend on the starting interval l, but only on the difference between j and l.

6.4.2. User Equilibrium assignment models

For congested networks, dynamic equilibrium assignment can be specified through fixed-point models by combining supply and demand models both for rigid and elastic demand profiles. For within-day dynamic systems the dependency of travel times on link flows (occupancies) introduces two feedback cycles: in addition to the path cost and flow cycle (typical of within-day static user-equilibrium models), in the dynamic case link flows depend on travel times (see Fig. 6.4.2).

In the continuous-flow case, user-equilibrium models can be formally stated as a fixed-point problem in travel times, costs and flows derived from the following system of non-linear equations:

$$\boldsymbol{t}^{J}(\tau) = \boldsymbol{t}^{J}(\boldsymbol{f}(\tau)) \tag{6.4.4a}$$

$$\boldsymbol{TT}^{J}(\tau) = \Gamma(\boldsymbol{t}^{J}(\tau'); \tau' \leq \tau) \tag{6.4.4b}$$

$$\boldsymbol{V}_{od}(\tau/\tau_o) = \beta_t \boldsymbol{TT}^{f}(\tau) + \boldsymbol{EC}(\tau) + \beta_e \boldsymbol{EDP}(\tau, \tau_o) + \beta_l \boldsymbol{LDP}(\tau, \tau_o) \tag{6.4.4c}$$

$$\boldsymbol{h}(\tau) = \sum_{\tau_o} \boldsymbol{P}\left(\boldsymbol{V}_{od}(\tau/\tau_o)\right) \cdot \boldsymbol{d}(\tau) \tag{6.4.4d}$$

$$\boldsymbol{f}(\tau) = \Phi\left(\boldsymbol{h}(\tau), \boldsymbol{t}^{f}(\tau); \tau' \leq \tau\right) \tag{6.4.4e}$$

Equation (6.4.4e) expresses the dependency of link flow vector at time τ, $\boldsymbol{f}(\tau)$, on the path flow vectors $\boldsymbol{h}$ and on link travel time vectors $\boldsymbol{t}$ in all previous time instants $\tau' < \tau$. This can be more explicitly stated, for instance, in terms of in-flows on the generic link a, as (see section 6.2.1.4):

$$u_a(\tau) = \sum\nolimits_k \delta_{ak} \cdot h_k(\tau - T^{b}_{a^k}(\tau)) \cdot \left(1 - \frac{dT^{b}_{a^k}(\tau)}{d\tau}\right)$$

where the dependency of link flow at τ on the travel times of all links of the network and in all previous instants τ' is embedded in the backward travel time, given by the recursive equation (6.2.16b) here restated for readers convenience:

$$T^{b}_{a^k_i}(\tau) = t^{b}_{a^k_{i-1}}(\tau) + t^{b}_{a^k_{i-2}}(\tau - t^{b}_{a^k_{i-1}}(\tau)) + t^{b}_{a^k_{i-3}}(\tau - t^{b}_{a^k_{i-1}}(\tau)) - t^{b}_{a^k_{i-2}}(\tau - t^{b}_{a^k_{i-3}}(\tau)) + \ldots$$
$$\ldots + t^{b}_{a^k_1}(\tau - t^{b}_{a^k_{i-1}}(\tau) - \ldots - t^{b}_{a^k_2}(\tau - t^{b}_{a^k_{i-1}}(\tau) - \ldots - t^{b}_{a^k_3}(\tau - t^{b}_{a^k_{i-1}}(\tau) - \ldots)))$$

A formal fixed-point specification in link flows of dynamic user-equilibrium continuos-flow models is the following:

$$\boldsymbol{f}^{*}(\tau) = \Phi\left(\sum_{\tau_o} \boldsymbol{P}\left(\beta_t \Gamma\left(\boldsymbol{t}(\boldsymbol{f}^{*}(\tau'))\right) + \boldsymbol{EC}\left(\boldsymbol{t}(\boldsymbol{f}^{*}(\tau'))\right) + \beta_e \boldsymbol{EDP}(\tau, \tau_o) + \beta_l \boldsymbol{LDP}(\tau, \tau_o)\right) \cdot \boldsymbol{d}_{\tau_o},\right.$$
$$\left. \boldsymbol{t}^{f}(\boldsymbol{f}^{*}(\tau')); \tau' < \tau \right)$$

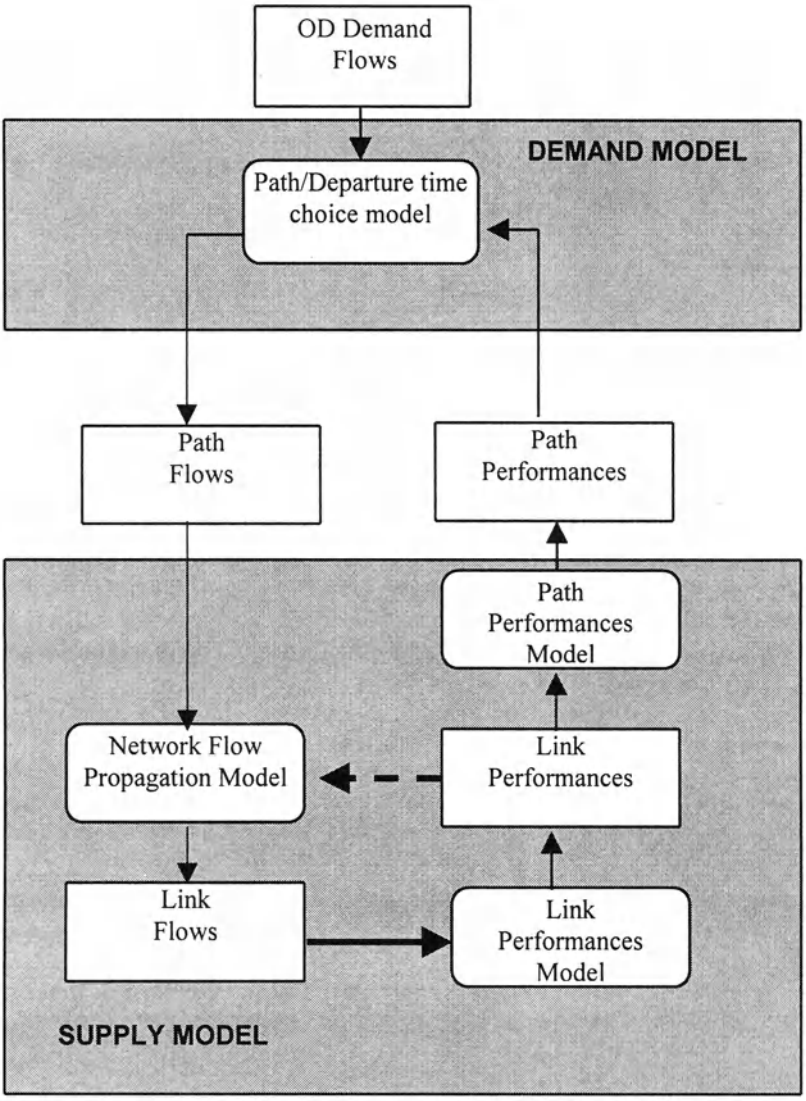


Fig. 6.4.2 Dynamic User Equilibrium Traffic Assignment.

Dynamic user equilibrium models can be deterministic or stochastic depending on the path-departure time choice model. Existence and uniqueness conditions for continuous-flow dynamic user equilibrium models are currently being studied (see the bibliographic note of this chapter).

In the discrete-flow case, the models can be formally formulated as follows:

$$\boldsymbol{t}^{f}(\tau_j) = \boldsymbol{t}^{f}(\boldsymbol{f}(\tau_j)) \tag{6.4.5a}$$

$$\boldsymbol{TT}^{f}(\tau_j) = \Gamma[\boldsymbol{t}^{f}(\tau_{j'})]; j' = 1 \ldots j] \tag{6.4.5b}$$

$$V_{od}(\tau_j / \tau_o) = \beta_t \boldsymbol{TT}^{f}(\tau_j) + \boldsymbol{EC}(\tau_j) + \beta_e \boldsymbol{EDP}(\tau_j, \tau_o) + \beta_l \boldsymbol{LDP}(\tau_j, \tau_o) \tag{6.4.5c}$$

$$\boldsymbol{h}(\tau_j) = \sum_{\tau_o} \boldsymbol{P}\left(V_{od}(\tau_j / \tau_o)\right) \cdot \boldsymbol{d}(\tau_o) \tag{6.4.5d}$$

$$\boldsymbol{f}(\tau_{j'}) = \Phi\left(\boldsymbol{h}(\tau_{j'}), \boldsymbol{t}^{f}(\tau_{j'}); j' = 1 \ldots j\right) \tag{6.4.5e}$$

Equation (6.4.5e) is the analogous of (6.4.1e) for the uncongested network case. It can be stated more explicitly as:

$$\boldsymbol{f}[j] = \sum_{l \le j} \Delta[l, j] \cdot \boldsymbol{h}[l] \tag{6.4.5f}$$

The difference with respect to the uncongested network is that, in this case, Δ is a function of links travel times $\boldsymbol{t}$ in all the previous intervals up to interval j:

$$\Delta[l, j] = \Delta[l, j]\left(\boldsymbol{t}^{f}(\tau_i); i = l \ldots j\right) \tag{6.4.5g}$$

A formal fixed point specification of a dynamic user equilibrium models is the following.

$$\boldsymbol{f}^{*}[\tau_j] = \sum_{\tau_o} \sum_{l=1 \ldots j} \Delta[l, j]\left(\boldsymbol{t}^{f}\left(\boldsymbol{f}^{*}[i]; i = l \ldots j\right)\right) \cdot$$

$$\boldsymbol{P}\left[\beta_t \Gamma\left(\boldsymbol{t}^{f}\left(\boldsymbol{f}^{*}[i]; i = l \ldots j\right)\right) + \boldsymbol{EC}\left(\boldsymbol{t}^{f}\left(\boldsymbol{f}^{*}[i]; i = l \ldots j\right)\right) + \beta_e \boldsymbol{EDP}(\tau_l, \tau_o) + \beta_l \boldsymbol{LDP}(\tau_l, \tau_o)\right] \cdot \boldsymbol{d}_{\tau_o}$$

Existence and uniqueness conditions for the fixed-point formulation have not been stated; however, in this case it is more difficult to arrive at general conditions, if possible at all, given the discreteness of time and packets.

6.4.3. Dynamic Process assignment models

Dynamic process models require further demand models simulating learning, or utility updating, and choice updating mechanisms as in the static case (see Fig. 6.4.3). These models can be seen as doubly-dynamic assignment models.

As in the static case, to formalize a dynamic process model we need to distinguish between *expected* (or *anticipated*) and *actual* path performance attributes at day *t*. The former are the attributes (e.g. the travel time on a give path) that users expect to encounter on the network at a given day *t*; the latter are what they actually experience. On the other hand, not all the users may reconsider their choices every day *t* due to inertia and/or habit.

In the discrete-flow case, let us consider for sake of simplicity, that path travel time is the only attribute updated from one day to the next and let:

$\boldsymbol{TT}^{f,t}_{exp}(\tau_j)$ be the (forward) travel time leaving at the representative time instant τ_j that users expect to experience at day *t*;

$\boldsymbol{TT}^{f,t}_{acl}(\tau_j)$ be the (forward) travel time leaving at the representative time instant τ_j that users actually experience at day *t*;

A Deterministic Dynamic Process model, based on simple exponential filters for travel times and choice updating models can be formally stated as follows:

$$\boldsymbol{t}^{f,t-1}(\tau_j) = \boldsymbol{t}^{f,t-1}(\boldsymbol{f}(\tau_j)) \tag{6.4.6a}$$

$$\boldsymbol{TT}^{f,t-1}_{acl}(\tau_j) = \Gamma(\boldsymbol{t}^{f,t-1}(\tau_{j'}); j'=1...j) \tag{6.4.6b}$$

$$\boldsymbol{TT}^{f,t}_{exp}(\tau_j) = \beta\, \boldsymbol{TT}^{f,t-1}_{acl}(\tau_j) + (1-\beta)\, \boldsymbol{TT}^{f,t-1}_{exp}(\tau_j) \tag{6.4.6c}$$

$$V^{t}_{od}(\tau_j/\tau_o) = \beta_t \boldsymbol{TT}^{f,t}_{exp}(\tau_j) + \boldsymbol{EC}(\tau_j) + \beta_e \boldsymbol{EDP}(\tau_j,\tau_o) + \beta_l \boldsymbol{LDP}(\tau_j,\tau_o) \tag{6.4.6d}$$

$$\boldsymbol{h}^{t}(\tau_j) = \alpha \sum_{\tau_o} \boldsymbol{P}\left(V^{t}_{od}(\tau_j/\tau_o)\right) \cdot \boldsymbol{d}(\tau_o) + (1-\alpha)\cdot \boldsymbol{h}^{t-1}(\tau_j) \tag{6.4.6e}$$

$$\boldsymbol{f}^{t}[j] = \Phi\left(\boldsymbol{h}^{t}(\tau_{j'}), \boldsymbol{t}^{f,t}(\tau_{j'}); j'=1...j\right) \tag{6.4.6f}$$

where β and α are respectively the weight given to the experience of the previous day *t-1* and the fraction of users reconsidering their choice (assumed here to be constant for each day *t*).

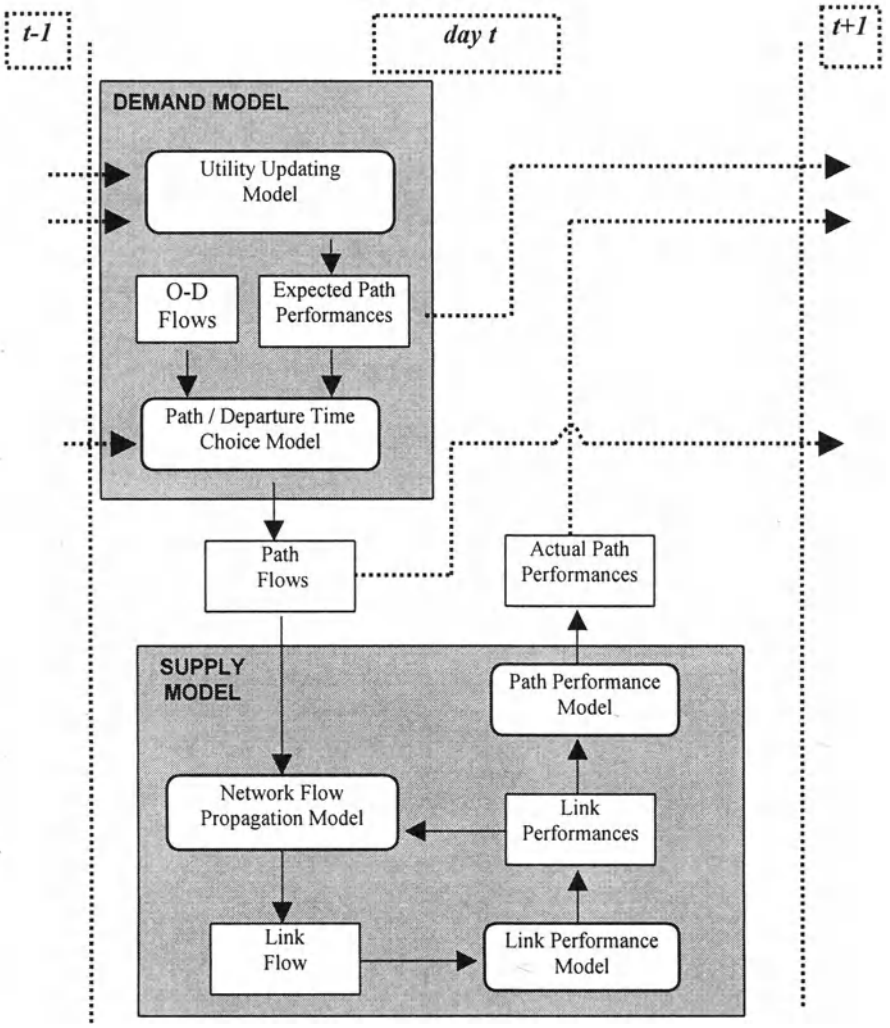


Fig. 6.4.3 Dynamic Process Assignment model (without information).

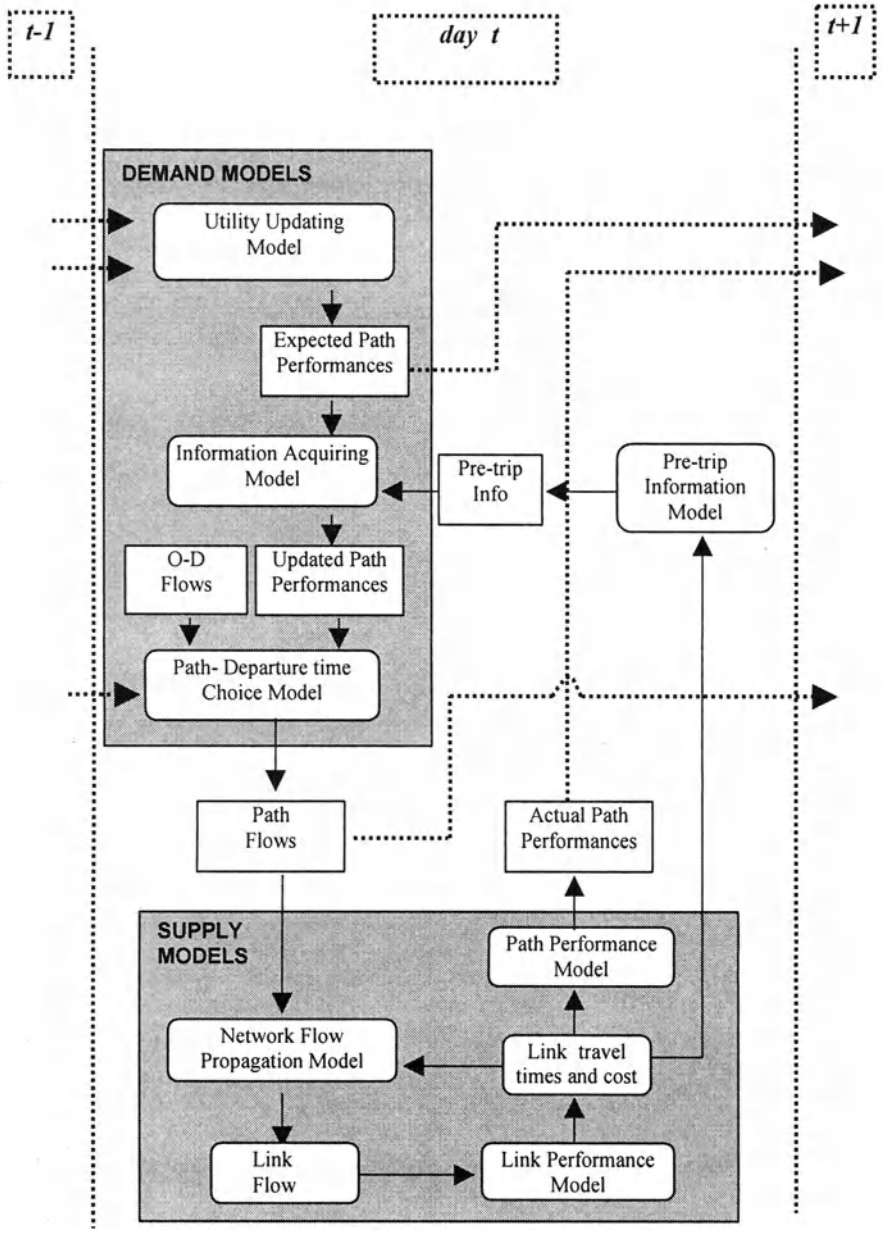


Fig. 6.4.4 Dynamic Process Assignment model (with Pre-trip information).

Note that given the mesoscopic nature of the model, individual packets updating models can be easily adopted. In this case, for instance, it is possible to update only the travel time experienced in the actual "yesterday" trip.

Dynamic process models for within-day dynamic systems can be further expanded to include real-time information available by (some) users. This is an expanding class of assignment models due to the growing interest in Advanced Traveler Information systems (ATIS). It is possible to distinguish two cases: information available only before starting the trip (i.e. *pre-trip information*) and information available during the journey (i.e. *en-route* or *while-trip information*). The former case requires other demand models to express the information acquiring process (see Fig. 6.4.4), the latter case requires as well demand models simulating compliance with prescriptive information at diversion nodes (see Fig. 6.4.5)

Dynamic process models can be deterministic or stochastic, as in the case of within-day static case, depending on the assumption made on the variables involved (average or deterministic variables or random variables). The full specification of these models requires assumptions on the type of information given and the information strategy, i.e. how information is related to actual system states (see Fig. 6.4.6). In general, several information strategies are possible: the ATIS can provide, for instance, *historical* information based on network performances in all previous time periods with similar characteristic (e.g. time of the day, day of the week, weather condition,...) or it can provide *real-time* information on current network conditions or it can predict what is going to happen on the network, i.e. *predictive* information. It is worth noting that predictive information is derived from prediction of future conditions, but these conditions are themselves affected by how users react to information they receive. In other words, there is a circular dependency between predictive information and network performance that can be seen again as a fixed-point problem. Furthermore, with respect to typology of the information provided these can be described as *descriptive* (i.e. travel or congestion phenomena) or *prescriptive* (i.e. route guidance or turning movements).

Due to the multiplicity of informative contents and the necessity to distinguish between users' categories (e.g. informed and non-informed, regular ad non-regular, etc...) it is not possible to give general formulation for dynamic assignment models with ATIS; for these reasons these models will not be described here.

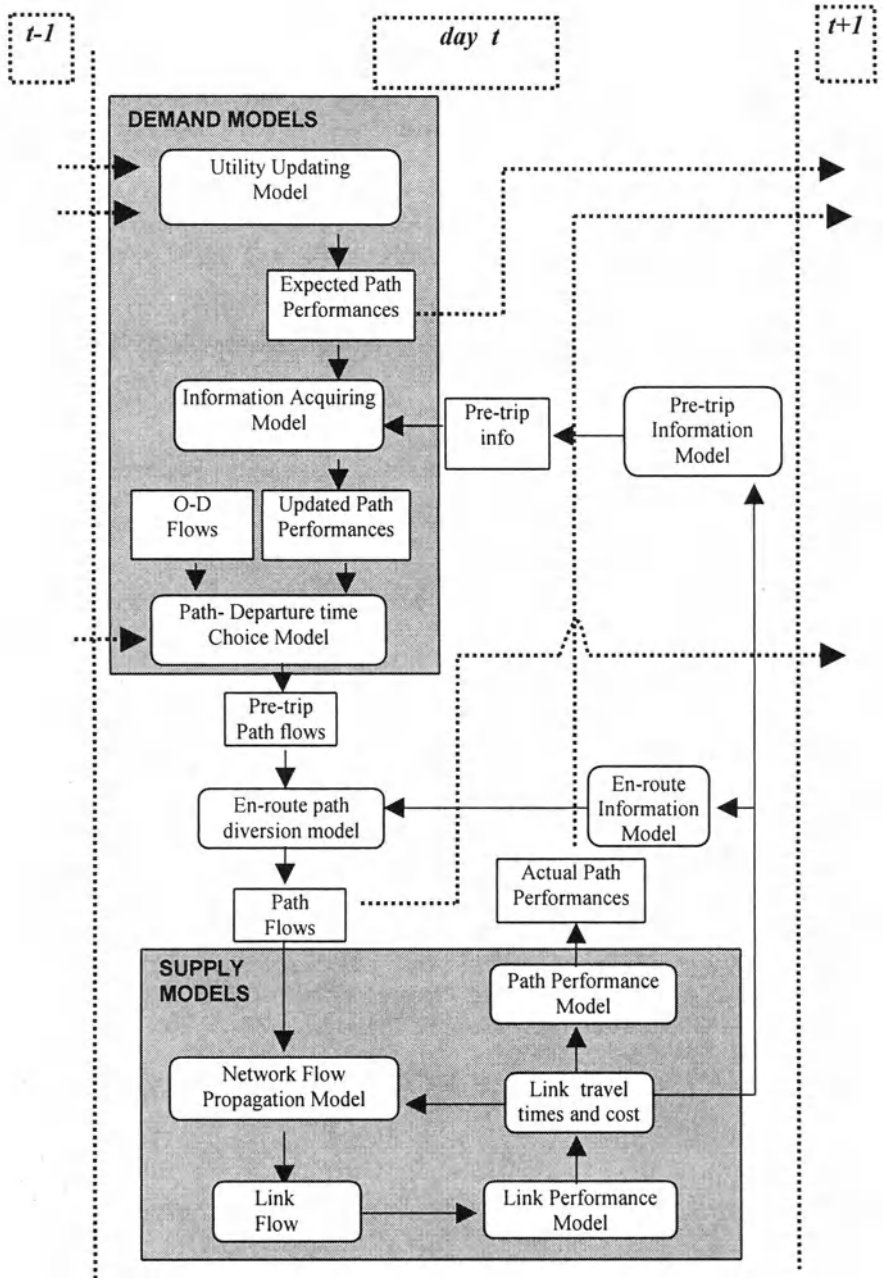


Fig. 6.4.5 Dynamic Process Assignment model (with Pre-trip/En-route information).

	Classification	Example
INFORMATION TYPOLOGY	Descriptive	"congestion ahead" "travel time to airport 5 min"
	Prescriptive	"turn left"
INFORMATION AVAILABILITY	Pre-trip	Information available via Internet, or Television
	En-route (or while-trip)	Variable Message Signs (VMS) or In-Vehicle Navigation Systems (IVNS)
INFORMATION TIME-DIMENSION	Historical	
	Real-time (or current)	
	Predictive (or self-consistent)	

Fig. 6.4.6 - Classification of information types.

6.5. Models for scheduled service systems(°)

Scheduled transportation services, such as those provided by airplanes, trains and buses, are considered discrete both in time and space as they can only be accessed at certain times and certain locations such as airports, rail stations, and bus stops. In a within-day dynamic context, it is possible to model explicitly supply, demand and demand-supply interactions for systems with scheduled services starting from the timetable. With respect to a given timetable, runs and lines can be defined (see Fig. 6.5.1). A *run, r*, represents a connection with a given time schedule (e.g. a given train connection), while a *line, ln*, as defined in Chapter 2, may be regarded as a set of runs of similar characteristics (e.g. stops, travel times, quality of services, etc.). Within-day dynamic models simulate explicitly supply and demand for runs rather than for lines, as was the case in static models for scheduled services systems described in previous chapters.

run	Line	Service type	Initial Station	Departure Time	Intermediate Stops	Terminal Station
1	AA	Intercity	A	9.30	-	D
2	BB	Regional	A	9.50	B/C	D
3	AA	Intercity	A	10.30	-	D
4	CC	Intercity	A	11.30	D	E

Fig. 6.5.1 Time schedule, runs and lines.

Dynamic models used to simulate within-day dynamic scheduled services systems differ according to a number of factors related to service characteristics. The main classification factors affecting dynamic models are *frequency, regularity* and *information* available to users.

Service *frequency* can be related directly to the frequency of the line in the reference period, i.e. the number of runs belonging to the line in such a period or, for overlapping lines, to the cumulative frequency i.e. the sum of the frequencies of all attractive lines connecting the O-D pair *od*.

The service *regularity* is a measure of how much the schedule is followed. Regularity, or rather its opposite, can be measured by several variables depending on the analysis purpose. If regularity is used to make assumptions on user behavior in line-based systems, such as buses and trains, deviations from the schedule should be related to the average headway of runs belonging to the same line.

(°)Agostino Nuzzolo is the co-author of this section.

Usually regular services are associated to low frequencies, typical of extra-urban systems such as (intercity) rail or air. On the other hand, irregular services generally correspond to high frequencies as in urban or metropolitan areas, e.g. bus or underground lines. In any case, *frequency* and *regularity* are continuous variables and their segmentation in "high" and "low" is conventional and somewhat arbitrary. In models, they correspond to different hypotheses on users' behavior and to different model systems. As such they are at the analyst's discretion.

Information on services can be available to the user pre-trip (i.e. at home) and/or en-route (i.e. at stops). In both cases the information can include waiting times, travel times and on-board occupancy. Static information on run schedule is traditionally available with timetables: Intelligent Transportation Systems (ITS) have expanded significantly the range of information available to the traveler, through Advanced Traveler Information Systems (ATIS), and improved the performances of transit services, through Advanced Public Transportation Control Systems (APTCS).

Different supply and demand models are used to simulate scheduled services systems depending on their different characteristics. In the case of *low frequencies* and *regular services*, supply is modeled through deterministic dynamic networks and users are assumed to have full information before starting their trip. They choose a specific run on the basis of the expected performance attributes, with models analogous to those assumed for modeling path choice on continuous service networks (see section 4.3.4.1).

On the other hand, supply models for *high frequencies* and *irregular services* are based on stochastic dynamic networks. Furthermore it is assumed that users may not have all information before starting their trip and, as described in section 4.3.4.2, they follow a mixed pre-trip/en-route choice behavior. It is commonly assumed that en-route choices occur at stops and are relative to the decision to board a particular run or to wait for another run of an attractive set. The choice of boarding stops is considered to be made before starting the trip, since it is not influenced by unknown events.

As usual, dynamic assignment models to scheduled services can be decomposed in supply, demand and supply/demand interaction models. A general scheme of intra-period dynamic assignment models to scheduled services is shown in Fig. 6.5.2.

In the following the two cases of low-frequency regular services and high-frequency irregular services will be addressed separately.

It should be noted that dynamic traffic assignment models for scheduled services is a newer and significantly less researched subject than DTA for continuous service (road) systems. The models described are thus somewhat less established than in the continuous case.

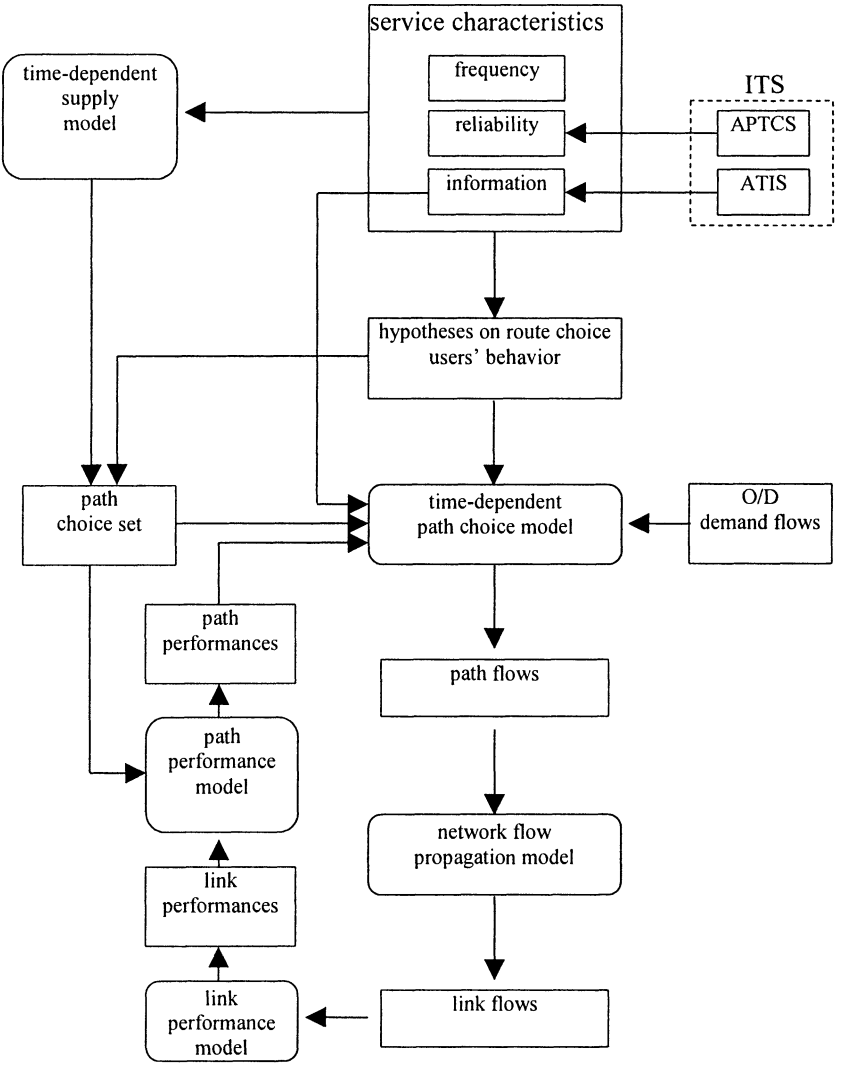


Fig. 6.5.2 – Schematic representation of within-day dynamic transit assignment models.

6.5.1. Models for regular low-frequency services

For regular low-frequency services, it is assumed that each run follows scheduled departure and arrival times, and that users have all relevant information before

starting their trips and choose access/egress terminals as well as the runs according to their desired arrival or departure times.

In the following sub-sections the within-day dynamic supply, demand and demand-supply interaction models will be discussed.

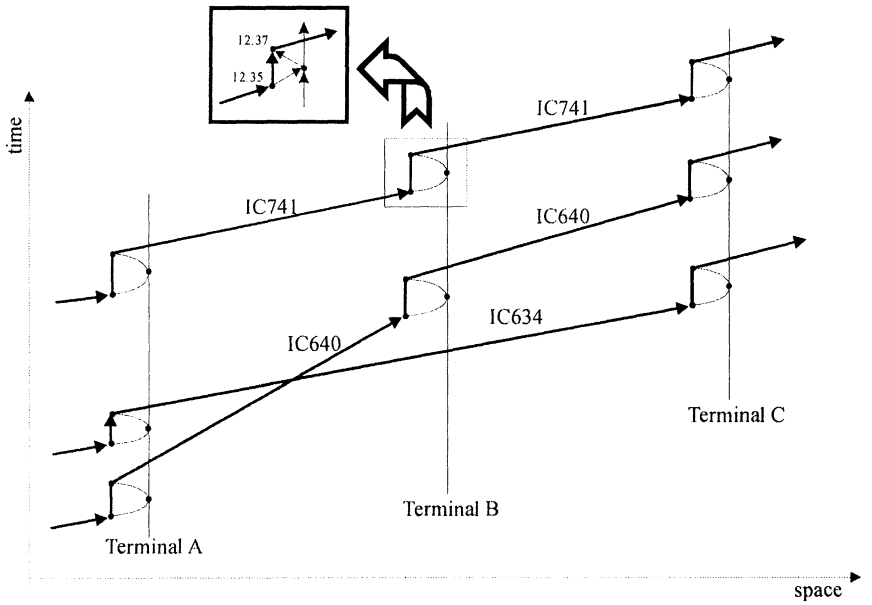
6.5.1.1. Supply models

In general, intra-period dynamic supply models of scheduled services consist of a network model (graph plus link performance and cost functions) and the relationships connecting path costs to link costs and link flows to path flows (network loading or flow propagation model). The main difference with continuous service (road) systems is in the graph model, which allows to adopt the convenient linear supply models introduced in Chapter 2 for static systems.

The model used for scheduled services is known as space-time or *diachronic* graph; in this graph some nodes can have an explicit time coordinate and therefore represent events taking place at a given instant. Each run can be described by means of a sub-graph (Fig. 6.5.3) whose nodes represent the arrival and departure times of the vehicles (trains, planes, buses) at the stations and whose links represent the travel from one station to another or the dwelling at a given station. Other nodes represent the arrival of the user at the station to board or alight from each run. These nodes are connected, through boarding and alighting links, to the nodes representing the arrival and departure of that run and by links representing the traveler's transfer from one run to another at the same stop. This set of nodes and links is usually defined as a run sub-graph.

Other sub-graphs are the temporal centroid graphs representing times and location of trips departure/arrival. To simulate the users' choices among different runs, or sequences of runs, it is necessary to introduce the desired departure times from the origin τ_o, or the desired arrival times at the destination, τ_d. Even if desired departure or arrival times are continuous variables, in applications discrete time intervals (e.g. five minutes long) are used. Possible desired departure or arrival times are represented as time centroid nodes having the same spatial coordinate of the zone centroids introduced in Chapters 1 and 2 and temporal coordinates given by representative time instants of the relative discrete time intervals (e.g. one node every five minutes). Nodes of the temporal centroid graph represent also the actual departure times from the origin to the boarding terminal or the actual arrival times at the destination from the alighting terminal. The links connecting the temporal centroid, representing the desired departure time of the generic user, to the temporal node representing the actual departure time from the origin to catch a given run represent the anticipation or delay of the actual departure with respect to the desired one (Fig. 6.5.4). Similarly the anticipation or delay for the actual arrival times with respect to the desired ones are represented by links in the temporal centroid graph.

The graph model for the overall system is usually completed with links representing access (egress) from (to) the centroids, with the relative travel times and costs. Fig. 6.5.4 shows a diachronic graph for desired departure time; similar graphs can be built for a desired arrival time.



TIMETABLE

run	Terminal A		Terminal B		Terminal C	
	arr.	par.	arr.	par.	arr.	par.
IC634	08.25	08.30	---	---	12.00	12.05
IC640	08.55	09.00	10.10	10.15	11.15	11.18
IC741	10.58	11.00	12.35	12.37	14.00	14.02

Fig. 6.5.3 Diachronic graph representation of scheduled services.

Diachronic graphs are very convenient since they exploit the intrinsically discrete service structure (the services being available only at certain time instants); this allows the use of very efficient network algorithms similar to those described for static continuous networks. Other models to represent regular services are based explicitly on timetable manipulations. These models are conceptually analogous to the graph representation, which is more consistent with the general approach to supply modeling followed throughout the book.

In a diachronic graph such as the one introduced, a trip is represented by a path k starting from the desired departure time on the temporal centroid sub-graph and ending at the arrival time at destination (see Fig. 6.5.4). Note that, unlike continuous

service graphs, desired departure time is uniquely associated to each path. The same sequence of runs for a different desired departure time correspond to a different path k'. Analogously a path k identifies uniquely the actual departure time (interval) τ_j.

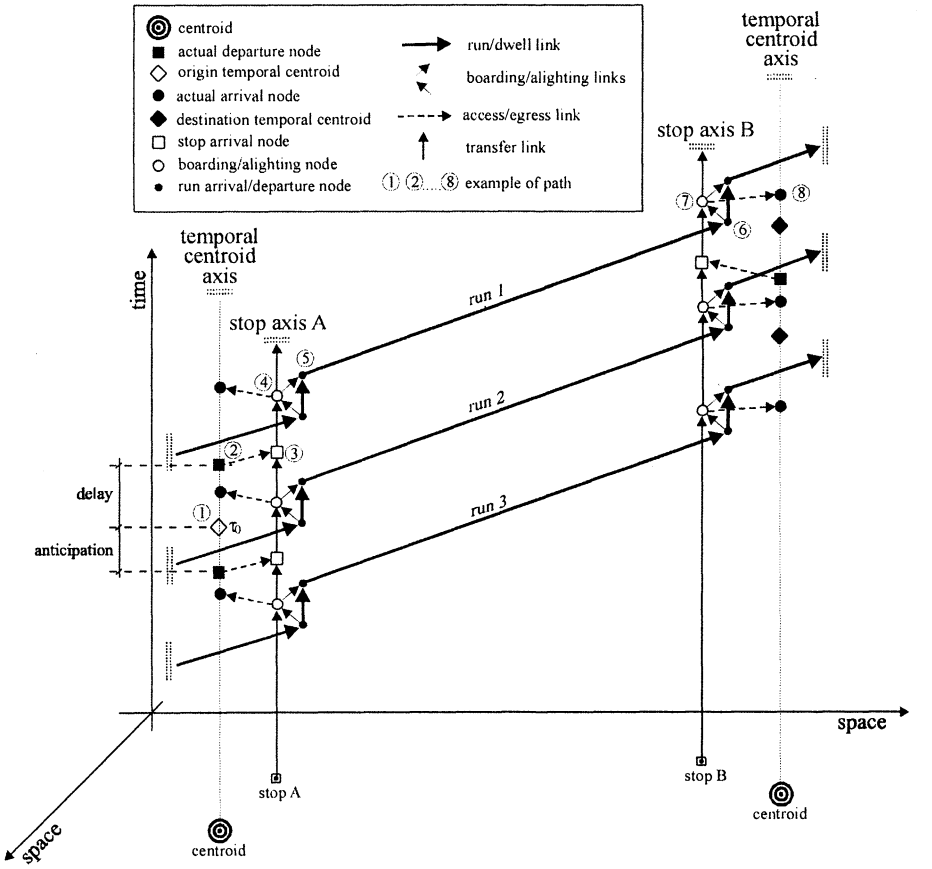


Fig. 6.5.4 Example of diachronic graph for low-frequency services.

For diachronic network models, performance variables, as well as their relationships with flows, are substantially analogous to those described for static models. As in Chapter 2, link users' performance or level of service attributes, r_{nl}, are variables expressing average values of individual attributes perceived by the users and associated to a given link. Examples of link attributes are monetary cost, access time, anticipation or delay with respect to scheduled departure, on board travel time, number of transfers, egress time, and so on. In the same way the *average generalized transportation cost*, or more synthetically the *link cost*, is the global

disutility associated to each link. The link cost, c_l, is a (dis)utility function, typically linear, of link performance attributes, underlying travel related choices and, more in particular, path choices:

$$c_l = \sum_i \beta_i \cdot r_{il}$$

Because of *congestion*, some performance attributes, and thus the average link cost, can depend on the number of users on the link and, possibly, on other links of the graph. Given the characteristics of the system (low-frequency and regularity, place booking, etc.), it is often assumed that link performances and costs are independent of flows, and supply is modeled as a non-congested network. In some cases, however, it can be appropriate to take into account congestion effects. This could be the case for regional buses or rail services, where not all passengers may have a seat and in the limit they may have difficulty to board some runs. Referring to on-board links (see Fig. 6.5.5), separable cost functions similar to those introduced in section 2.3.2.2 may be used to represent discomfort (equation 2.3.20) and on-board travel time (equation 2.3.15). Penalty functions can be adopted to represent possibility of not being able to board a given run due to overcrowding.

Note that early or late schedule penalties, $EAP_k(\tau, \tau_d)$ or $EDP_k(\tau, \tau_o)$ and $LAP_k(\tau, \tau_d)$ or $LDP_k(\tau, \tau_o)$ introduced in section 6.3 for continuous service dynamic demand models can be obtained as additive link costs related to the links in the temporal centroid graph connecting the node corresponding to the desired departure (arrival) time with the node corresponding to the actual departure (arrival), see Fig. 6.5.4.

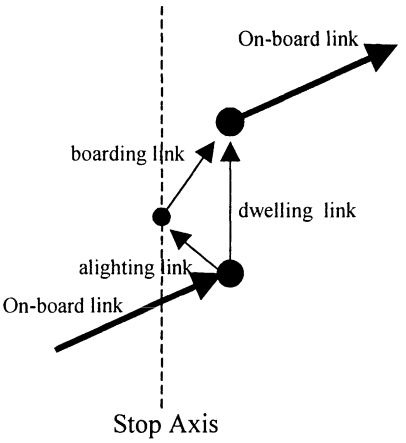


Fig. 6.5.5 Link classification at stops.

Performance attributes and generalized transportation cost (disutility) can be extended from links to paths. The average generalized transportation cost g_k of a

generic path k is defined as a scalar quantity homogenizing the different performance attributes perceived by the users for the whole trip. As in Chapter 2, the path cost in the most general case is made up of two parts: link-wise additive cost, g_k^{ADD}, and non-additive cost, g_k^{NA}, assuming that they are homogeneous:

$$g_k = g_k^{ADD} + g_k^{NA} = \sum_l \delta_{lk} c_l + g_k^{NA} \tag{6.5.1}$$

or in matrix terms:

$$\boldsymbol{g} = \Delta^T \boldsymbol{c} + \boldsymbol{g}^{NA} \tag{6.5.2}$$

where Δ expresses the link-path incidence matrix. The price structure can be non linear with respect to distance (e.g. based on the O/D pair, independently on the run or sequence of runs followed), thus requiring the introduction of non-additive costs.

The average number of users (in a time unit) following path k is called the *path flow* h_k. The *link flow*, f_l, represents the average number of travelers using the link. Thus flow on a link representing a connection between two successive stops of a particular run is the average number of users using that service segment. To adopt a terminology and notation consistent with within-day static models, the number of users on a link or following a path in the diachronic network has been referred to as *flow*, even though it is conceptually and dimensionally a number rather than a rate for time unit.

In within-day dynamic supply models for scheduled services, link flows can be obtained by summing flows on all paths including that link. This leads to a linear network loading model as in the within-day static case:

$$f_l = \sum_k \delta_{lk} h_k \tag{6.5.3}$$

$$\boldsymbol{f} = \Delta \boldsymbol{h} \tag{6.5.4}$$

6.5.1.2. Demand models

Demand models used in dynamic assignment for low-frequency regular scheduled service networks are analogous to those described for discrete-time models of continuous services and express the relationship between path flows and path costs.

The users' flow following a path k connecting the O-D pair od and starting the trip in each interval $[j]$ can be obtained with *elastic demand profile* models, also simulating the departure interval choice as a function of the desired arrival time, τ_d, or the desired departure time, τ_o. In this case there is no need to model departure time choice separately from path choice since the former is implicitly included in each path alternative.

Path choice models for single class-single service give the probability $p_{od,k}(\tau_j/\tau_o)$ of choosing path k and related actual departure time τ_j, given O-D pair od and desired departure time τ_o (or alternatively desired arrival time τ_d). Pre-trip path choice models assume that users choose the path minimizing the perceived disutility, taking into account several attributes such as access and egress times and costs, travel time, number of transfers, monetary cost, comfort and early/late schedule delay. These attributes are typically homogenized in the path cost variable introduced in the previous section. Other attributes (e.g. socio-economic variables) can be included in a term V_{ok}.

Most models proposed in the literature to simulate path choice also simulate choice set following a selective approach, see section 4.3.4. In particular it is assumed that only some of the topologically feasible paths belong to the choice set. Paths are selected following dominance rules such as:

- runs leaving before and arriving after with respect to other runs in the choice set are not included in the set;
- paths must satisfy criteria relative to maximum number of transfer, maximum time of transfer, maximum travel time and so on.

The global systematic utility of a given path k can thus be expressed as:

$$V_{od,k}(\tau_j/\tau_o) = g_k(\tau_j/\tau_o) + V_{ok} \tag{6.5.5}$$

Note that in equation (6.5.5) the departure time τ_j of the first run and the desired departure time τ_o, both univocally associated to path k, have been explained in analogy with continuous service models.

A Logit specification of the path choice model for desired departure time τ_o at destination is:

$$p_{od,k}(\tau_j / \tau_o) = \frac{exp\left(V_{od,k}(\tau_j / \tau_o)\right)}{\sum_{k'} exp\left(V_{od,k'}(\tau_j / \tau_o)\right)} \tag{6.5.6}$$

If there are several service types (e.g. intercity or regional) and classes (e.g. first/second class) the interdependence of choice dimensions can be accounted for by assuming a positive correlation among random residuals of perceived utilities of paths sharing service type, class, etc. In this case a multi-level hierarchical Logit path choice model could be adopted.

The average flow h_k on path k can thus be expressed as:

$$h_k = d_{od}(\tau_o) \cdot p_{od,k}(\tau_j / \tau_o) \tag{6.5.7}$$

Note that equation (6.5.7) is the equivalent of equation (6.2.40) for continuous-services continuous-flow models.

6.5.1.3. Demand-supply interaction models

Given the supply and demand models described in the previous sub-sections, within-day dynamic assignment models to regular low-frequency scheduled service networks reduce to within-day static assignment on a diachronic network. Also in this case it is possible to distinguish between Uncongested Network, User Equilibrium and Dynamic Process assignment models.

Since paths correspond to composite choice alternatives including departure time/ access-egress terminals/ runs, only random utility models have been adopted and calibrated, thus giving rise to stochastic assignment models (SUN, SUE etc.).

The general theoretical results on existence and uniqueness of solutions described in Chapter 5 can be applied to this case and will not be repeated here.

6.5.2. Models for irregular high-frequency services

For irregular high-frequency services, the complexity of the real system increases considerably with respect to both users' behaviour and performance variables. Different within-day dynamic models can be specified for these systems under different assumptions. In this section one such models will be described stressing, once more, that this area is very little researched and more theoretical developments and applications are expected in the future.

In this model, users are assumed to make their choices at different times during their trips. The choice of the first boarding stop and the attractive line set is made before the trip begins (pre-trip choice). During the trip users choose the actual runs at stops adapting to the actual succession of run arrivals and to information given (if any) about waiting times. It is further assumed that because of the high-frequency and the irregularity of services, the actual departure time from the origin is equal to the desired departure time, so they arrive at stops independent of run departure times. Thus if τ_o is the (desired) departure time from the origin and $t_{a,os}$ the access time to stop s, arrives at the stop at the absolute time $\tau_{so}=\tau_o + t_{a,os}$.

In the following subsections the supply, demand and demand-supply interaction models consistent with the above assumptions will be described.

6.5.2.1. Supply models

The diachronic network model described in section 6.5.1, with some differences, can be adopted also in the case of irregular services. Due to the irregularity of services, the actual arrival and departure times of each run at day t can be different from the scheduled ones and from those in other days. This can be represented by a vector of a random variables $\boldsymbol{b}$ whose elements are the arrival time $b_{ar,s}$ and the departure time $b_{pr,s}$ of each run r at each stop s. In the following $\boldsymbol{b}^t$ indicates a realisation of vector $\boldsymbol{b}$ and G^t is the relative diachronic graph (see Fig. 6.5.4). The equations (6.5.1), (6.5.2), (6.5.3) and (6.5.4), expressing the relationships between path costs and flows with link cost and flows can be still used if a link-path incidence matrix Δ^t relative to graph G^t, is defined.

It is usually assumed that the mean of random variables $b_{a,rs}$ and $b_{pr,s}$ coincide with the scheduled arrival and departure times.

The vector $\boldsymbol{b}$ is related to another vector $\boldsymbol{y}$ whose components, represent the running time of run r on running link l, y_{rl}, and the dwelling time of run r at stop s (dwelling link), y_{rs}. Because of the irregularity of the services also $\boldsymbol{y}$ can be modelled as a vector of random variables. The components of two vectors $\boldsymbol{b}$ and $\boldsymbol{y}$ are related through the following recursive equations:

$$b_{a,rs}=b_{p,r(s-1)}+y_{r,l} \qquad l\equiv((s-1),s)$$

$$b_{p,rs}=b_{a,rs}+y_{r,s}$$

Thus, given the initial departure time of run r, from a given vector $\boldsymbol{y}^t$ it is possible to generate a vector $\boldsymbol{b}^t$ and vice-versa. In applications the random vector $\boldsymbol{y}$ is often modelled from empirical observations. One of the model proposed is a MultiVariate Normal (MVN) with mean $\bar{y}$ (scheduled running and dwelling times) and a variance-covariance matrix Σ_y whose elements can take into account, though implicitly , several circulation phenomena such as:

- the propagation of delays between successive sections of the same line,

$$\mathrm{cov}(y_{r,l-1}, y_{rl})>0$$

- the persistence of perturbation factors on a given line section,

$$\mathrm{cov}(y_{r,l}, y_{r+1,l})>0$$

- the reduction in dwelling time due to a longer dwelling time of the previous run at the same stop,

$$\mathrm{cov}(y_{r-1,s}, y_{rs})<0$$

From the algorithmic point of view a configuration G^t of the diachronic network can be generated by sampling a vector $\boldsymbol{b}^t$ or $\boldsymbol{y}^t$ from the multivariate distribution assumed for $\boldsymbol{b}$ or $\boldsymbol{y}$. For example, the Monte-Carlo method with a Cholesky factorialization of the matrix Σ_y can be used if $\boldsymbol{y}$ is assumed distributed as a $MVN(\bar{y}, \Sigma_y)$.

In any case the resulting vector $\boldsymbol{y}^t$ must be modified to satisfy some feasibility rules such as the congruence of generated times with the allowed speeds for transit vehicles, the absence of overtaking between successive runs, and so on.

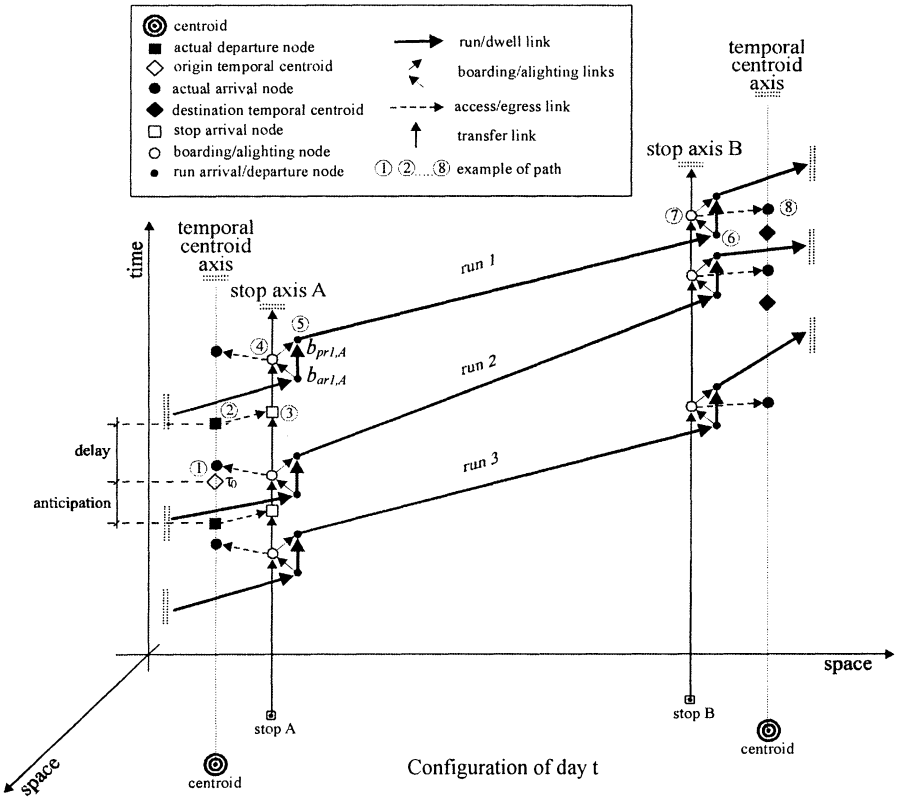


Fig. 6.5.6 Example of diachronic graph for high frequency irregular services.

6.5.2.2. Demand models

Generally for the same origin temporal centroid several different boarding stops, s, can be reached and many runs are available (see Fig. 6.5.6). Thus path choice on a realisation G' of the diachronic network implies choice of access stop and choice of the run(s) leading the user to the destination.

Path choice models give the probability $p_{od}[r,s|\tau_o]$ of choosing path including run r at boarding stop s, given the O-D pair od and the origin desired departure time τ_o (or the arrival time at stop s, $\tau_{s,o}$). Given the different choice behaviour assumed for pre-trip choices (stop s) and en-route choices (run r), this probability can be expressed as:

$$p_{od}[r,s/\tau_o] = p_{od}[r/s, \tau_{s,o}]\, p_{od}[s/\tau_o] \qquad (6.5.8)$$

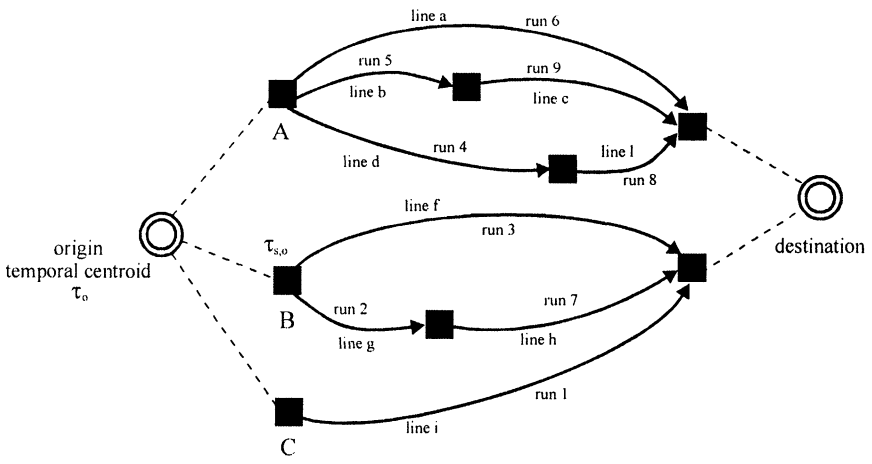


Fig. 6.5.7 Example of path choice set.

This is the product of the probability of choosing run r at stop s, given the arrival time $\tau_{s,o}$, by the probability of choosing stop s, given the desired origin departure time τ_o. In the following the index od, when not reported, is understood.

For modelling choice probabilities in equation (6.5.8), given the irregularity of services, some further assumptions have to be made on available information and the related choice set.

If information about waiting times is available at stops, the user can consider as choice alternatives runs of different lines according to their arrival times in any particular day t. Thus for users leaving origin "o" for destination "d" at time τ_o, arriving at stop s (where a user information system on run waiting or arrival times operates) at a time $\tau_{s,o}$ and finding a supply configuration $\boldsymbol{b}^t$, an initial choice set of runs $K^s[\tau_{s,o}, \boldsymbol{b}^t]$ may be defined. This set is specified by line runs connecting stop s directly or indirectly to destination d and satisfying some feasibility rules, such as:

- the set includes the first run of each line leaving after user arrival at stop at time $\tau_{s,o}$;
- the runs are not dominated (i.e. there are no runs leaving before and arriving after with respect to other runs of the choice set);
- the runs satisfy some criteria as the maximum number of transfers, maximum time of transfer, maximum travel time, etc.

The set $K^s[\tau_{s,o}, \boldsymbol{b}^t]$ depends on user arrival time at the same stop $\tau_{s,o}$, since different runs can be accessible to users for different arrival times; it depends also on the system configuration $\boldsymbol{b}^t$, since for the same arrival time on different days, different choice sets maybe available, due to system irregularities.

Furthermore, in case of oversaturation of arriving runs, the set can be modified while the user waits at the stop. When a run of a specific line included in $K^s[\tau_{s,o}, \boldsymbol{b}^t]$

arrives and has no available places, the user can decide to extend the choice set, introducing the next run of the same line.

For example, with reference to Fig. 6.5.8, for a configuration $\boldsymbol{b}^t$ and a user arriving at τ_1, the run choice set consists of run 1 of line *b*, run 2 of line *a* and run 1 of line *c*. This set will differ if the user arrives in τ_2 or if he arrives in τ_1 of day $t+1$ finding a different supply configuration $\boldsymbol{b}^{t+1}$. In the latter case, if there is congestion (e.g. on run *b*1) the choice set may be extended to run 2 of line *b*.

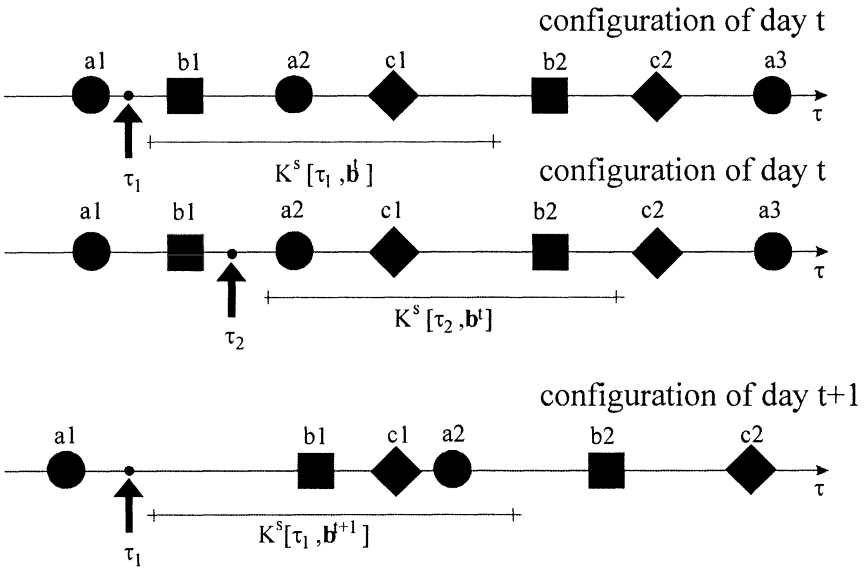


Fig. 6.5.8 Dependence of run choice set on configuration $\boldsymbol{b}^t$ and arrival time $\tau_{s,o}$.

The choice set may change while the user waits at the stop, not only because of congestion, but also because for each arrival, if not boarded, the corresponding run is eliminated from the set. This point will be clarified below.

A set of arrival times for the runs belonging to $K^s[\tau_{s,o}, \boldsymbol{b}^t]$ can be associated with each choice set $K^s[\tau_{s,o}, \boldsymbol{b}^t]$for any arrival time τ^+ of run r^+. In the following $K^s[\tau^+, \boldsymbol{b}^t]$ denotes the generic set available at time $\tau^+ > \tau_{s,o}$ of arrival of run r^+ at the stop, with respect to which the user makes his/her choice.

A sequential mechanism can be assumed to simulate run choice. When a run r^+ of the path choice set $K^s[\tau^+, \boldsymbol{b}^t]$ arrives at time $\tau^+ > \tau_{s,o}$, the user chooses in an intelligent adaptive way, to get on r^+ if the perceived utility U_{r+} is greater than the utility U_{r^*} of all other runs $r^* \in K^s[\tau^+, \boldsymbol{b}^t]$ yet to arrive. In formal terms it follows:

$$p_{od}[r^+/s,\tau^+] = Prob[U_{r+} > U_{r^*}] \quad \forall r^* \neq r^+ \quad with \tau^* > \tau^+ \qquad r^+ and\ r^* \in K^s[\tau^+, \underline{b}^t] \qquad (6.5.9a)$$

As usual perceived utilities can be expressed as the sum of a systematic utility, expressed a linear combination of attributes, and a random residual. A possible specification is:

$$U_{r+} = V_{r+} + \varepsilon_{r+} = \beta_{CFW} CFW_{r+} + \beta_b Tb_{r+} + \beta_c Tc_{r+} + \beta_{CFB} CFB_{r+} + \ldots$$
$$\ldots + \beta_n Nn_{r+} + \beta_p Tp_{r+} + \varepsilon_{r+} \qquad (6.5.9b)$$

$$U_{r*} = V_{r*} + \varepsilon_{r*} = \beta_{CFW} CFW_{r*} + \beta_b Tb_{r*} + \beta_c Tc_{r*} + \beta_{CFB} CFB_{r*} + \ldots$$
$$\ldots + \beta_n Nn_{r*} + \varepsilon_{r*} \qquad (6.5.9c)$$

where:

CFW_{r+}, CFW_{r*} are the boarding comfort (function of on-board crowding at stop);
Tw_{r*} is the waiting time (equal to the difference between the arrival time of run r^+ and the arrival time of run $r*$, provided by information systems);
Tb_{r+} and Tb_{r*} are on-board times;
Tc_{r+} and Tc_{r*} are transfer times;
Nn_{r+} and Nn_{r*} are the number of transfers;
CFB_{r+}, CFB_{r*} are the "route" on-board comfort (function of on-board crowding degrees in the following links);
Tp_{r+} is the time already spent at stop (equal to the difference between arrival time of run r^+ and the user arrival time τ_s at stop) simulating a possible "impatience effect" (β_p>0).

Note that in this model users cannot make their definitive choice upon arrival at stop at time τ_{so}, even if full information about waiting times is available, because boarding comfort degrees CFW are not known. Of course, if the user does not choose run r^+, the choice is reconsidered when the subsequent run arrives and so on (sequential run choice behaviour). Other more or less complex choice mechanisms can be assumed.

If it is assumed that random residuals ε in equation (6.5.9a) are i.i.d. Gumbel distributed, the choice probability $p_{od}[r^+/s, \tau^+]$ at time τ^+ of the arriving run r^+, conditional on not choosing previous runs and relative to the choice set $K^s[\tau^+, \boldsymbol{b}^t]$, can be expressed by a logit model:

$$p_{od}\left[r^+/s, \tau^+\right] = \frac{\exp\left(V_{r^+}\right)}{\sum_{r \in K^s\left[\tau^+, b^t\right]} \exp\left(V_r\right)} \qquad (6.5.10)$$

The total probability of choosing a given run r, can be expressed as the product of the conditional probability (6.5.10) and the probability of not having chosen any previous run r belonging to the choice set $K_s[\tau', \boldsymbol{b}^l]$:

$$p_{od}\left[r/s,\tau_{s,o}\right]=\prod_{r^-=1\ldots r-1}\left(1-p_{od}\left[r^-/s,\tau^-\right]\right)\cdot p[r/s,\tau] \qquad (6.5.11)$$

where each conditional probability depend on the arrival time $\tau_{s,o}$ and may be computed through equations (6.5.9) and (6.5.10).

The probability of choosing the boarding stop s, $p_{od}[s/\tau_o]$, can be specified with a different model referring to a choice set of boarding stops, S_{od}, that can be specified following different rules (e.g. by considering all stops within a certain distance from the origin). A perceived utility $U_s(\tau_o)$, can be associated to each stop in the choice set:

$$U_s(\tau_o)=V_s(\tau_o)+\varepsilon_s=\boldsymbol{\beta}^T\cdot\boldsymbol{X}_s+\beta_H\cdot H_s+\varepsilon_s \qquad (6.5.12)$$

where $\boldsymbol{\beta}^T$ is the vector of the model parameters, $\boldsymbol{X}_s$ is a vector of stop-specific attributes (e.g. access time, presence of shops, etc.), and H_s is an "inclusive utility" expressing the average utility associated to all runs available at stop s. To model the inclusive utility further assumptions have to be made on how travellers acquire and process information about the system performances. This model is strictly connected to the approach followed to simulate demand-supply interactions. One possible specification of H_s is based on the frequencies of the lines available at each stop and belonging to a feasible path on the line graph. This model is justified by the hypothesis of the lack of regularity (and information) and high frequencies of the system. Assuming a logit path choice model among the lines ln belonging to a set $Ln_s(o,d)$ of lines available at s given the O-D pair od, the inclusive utility is proportional to the logsum variable H_s:

$$H_s=ln\sum_{ln\in Ln_s(o,d)}exp\left(V_{ln,od}\right)$$

with $V_{ln,od}$ depending on average (scheduled) level of service attributes related to the line ln and given by:

$$V_{\mathrm{ln}}=\beta_w Tw_{\mathrm{ln}}+\beta_b Tb_{\mathrm{ln}}+\beta_c Tc_{\mathrm{ln}}+\beta_n Nn$$

where the symbols have the same interpretation as in equation (6.5.9) but the coefficients are in principle different since they represent different choice mechanism. Alternatively the average cost of the minimum hyper-path connecting s to the destination d can be associated with each stop s. This model has the advantage

of exploiting all the theoretical results and the computational algorithms described in Chapters 5 and 7. In this case it would result $H_s \equiv x_{sd}^{\min}$.

Using a Logit model, the stop choice probability can be expressed as:

$$p_{od}[s/\tau_o] = \frac{exp(V_s(\tau_o))}{\sum_{s' \in S_{od}} exp(V_{s'}(\tau_o))} \tag{6.5.13}$$

Thus the total choice probability of a path k represented by departure time τ_o, boarding stop s and run r (6.5.8) can be obtained through expressions (6.5.10), (6.5.11) and (6.5.13).

Finally, the average path flow h_k can be expressed as:

$$h_k = d_{od}(\tau_o) \cdot p_{od}[r, s/\tau_o] = d_{od}(\tau_o) \cdot p_{od}[r, s/\tau_o] \cdot p_{od}[s/\tau_o] \qquad k \equiv (\tau_o, s, r)$$

6.5.2.3. Demand-supply interaction models

Given the irregularity of the system and the assumptions made on users behaviour, especially at stops, demand-supply interactions should be consistently modelled through Stochastic Dynamic Process (SDP) model. In fact, service irregularities, represented by random vectors $\boldsymbol{b}$ and $\boldsymbol{y}$, are simulated through a stochastic supply model. On the other hand user choices at day t can be assumed as independent random variables following a multinomial distribution with path choice probabilities given by equation (6.5.8). Fig. 6.5.10 shows the number of users on the same section of the same run simulated in successive days for a real-size urban transit network under severe irregularity conditions.

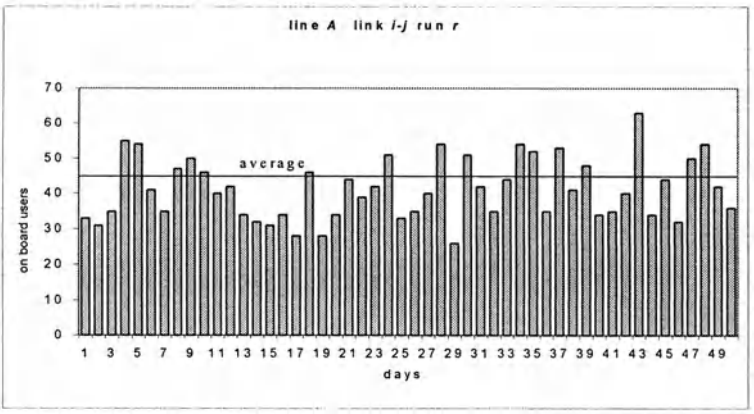


Fig. 6.5.10 Example of loads on the same section of the same run in different days.

The type of *SDP* model depends on a number of assumptions. The first factors are the assumptions made about users learning (cost-updating) mechanisms. If it is assumed that their pre-trip choices are based on average line attributes, see section 6.5.2.2, stop choice probabilities $p_{od}[s/\tau_o]$ do not change over successive days, while run choice probabilities are affected by random events occurring at each day t but do not depend on previous days. Under these assumptions the stochastic process is a renewal process, i.e. joint distribution probabilities of the variables describing the system state are independent from the states occupied in previous days. This assumption is reasonable for uncongested systems where explicit utility updating mechanisms can be ignored and users base their choices on line frequencies due to the unreliability of the timetable.

Matters are further complicated by congestion effects. Given the randomness of the system, congestion levels vary over successive days. If users are assumed to choose the boarding stop on the basis of uncongested attributes (as it's typical of irregular users), congestion plays a role only in run choices at stops and the stochastic process is still a renewal one. Otherwise (regular) users base their pre-trip choices on expected congestion levels resulting from their previous experience. In this case an utility updating filter similar to the ones described in section 5.8 has to be introduced and the process becomes a Markovian one.

Reference Notes

Although Dynamic Traffic Assignment is a relatively new research subject, a wide body of literature has been produced over the last 15 years (and only some of them are quoted here).

The first to propose DTA as a research subject of its own in a form similar to the present formulations were Ben Akiva et. al. (1984). The framework adopted in this chapter to present supply, demand, and supply interaction models, is original.

Continuous-flow models were first investigated by the scientific community. These models were adopted in the seminal work of Merchant and Nemhauser (1978) addressing system-optimal DTA with a single origin. The first to identify the Dynamic Network Loading model as a component of any DTA model were Cascetta and Cantarella (1991). The continuous-flow supply model described is based on the work of Friesz et al. (1989), who introduced the travel time link flow propagation model and equivalent conditions for respect of the FIFO rule. Recently more general equivalent condition for FIFO have been stated by Chabini and Kachani (1999), who have also investigated the properties of uniqueness and existence of the continuous-flow single-link network DNL problem. Some heuristic algorithms have been proposed in the literature for solving the supply model for general networks (Wu et al.; 1995, Astarita, 1996; Xu et al., 1994)

In the literature there are several papers proposing discrete flow supply models, both at the mesoscopic level (Cascetta and Cantarella, 1991; Jayakrishnan and Mahamassani., 1994; Ben Akiva et al., 1997; Cantarella et al., 1999), and at the

microscopic level (Yang and Koutsopoulos, 1996). In general it can be said that little or no efforts has been made to propose a general formulation of discrete flow models as well as to investigate their theoretical properties as for continuous models. Under this respect the proposed general framework for discrete flow models is original.

Demand models of departure-time choice were first proposed by Abkowitz (1981) and Small (1982); a joint departure time-path choice model for urban networks was proposed by Cascetta , Nuzzolo and Biggiero (1992). More complex departure-time and path switching models were proposed by Mahamassani and Liu (1999).

Most models proposed in the literature for demand-supply interactions are User Equilibrium models both deterministic and stochastic.

Papers on continuous flow models usually proposed ways to extend to time-varying demand and link flows the Deterministic (Wardrop's) User Equilibrium equivalent formulations (i.e. optimization or variational inequalities). Among these, the papers of Boyce et al. (1991), Janson (1989), Vytoulkas (1990), Friesz et al. (1993), Wie et al. (1990) and the book by Ran and Boyce (1994) can be referred to. Stochastic User Equilibrium has been formulated as a fixed-point problem by Daganzo (1983) and Cantarella (1997); however the general formulation of continuous and discrete flow within-day dynamic, fixed point models is original. Examples of dynamic process models are those proposed by Cascetta and Cantarella (1991), Cantarella et al. (1999), Jayakrishnan et al. (1994), Jha et al. (1998).

Dynamic assignment for transit, or other scheduled, services received considerably less attention in the literature. The idea to represent the schedule through a diachronic network can be credited to Nuzzolo and Russo. (1996). Some examples of dynamic assignment models for low-frequency regular services can be found in Cascetta et al. (1996) and Nuzzolo et al. (2000), for multiple-service rail networks, and in Cascetta and Papola (2000) for multimodal bus and rail networks. Dynamic assignment for irregular scheduled services is a still newer area. The papers by Hickman and Wilson (1995), Hickman and Bernstein (1997) and Nuzzolo et al. (1999) are among the few studying these models.

7 ALGORITHMS FOR TRAFFIC ASSIGNMENT TO TRANSPORTATION NETWORKS(°)

7.1. Introduction

In Chapter 5 several (within-day static) assignment models were formulated under for various assumptions on users' behavior and network congestion. Computing link flows and other relevant variables resulting from assignment is computationally for real size networks with thousands of nodes and tens of thousands of links and intensive requires efficient algorithms. This chapter describes the theoretical foundations and the structure of some of the simplest algorithms for solving (within-day) static assignment models (algorithms for within-day dynamic assignment presented in Chapter 6 are still at a reasearch stage). The main emphasis is on presenting simple and effective solution approaches for assignment to large-scale networks, rather than providing an exhaustive analysis of the many existing algorithms.

Section 7.2 describes the structure of shortest path algorithms used within many assignment algorithms.

Section 7.3 describes the algorithms for stochastic and deterministic uncongested network assignment. These algorithms will also be used as elements of algorithms for equilibrium assignment to congested networks. Uncongested network assignment algorithms can easily be extended to compute the satisfaction function (described in section 3.5) as a function of link costs. The satisfaction function is also relevant for elastic demand assignment algorithms (section 7.6) since its value is an attribute of the demand functions (specified by the model system described in Chapter 4). Furthermore, satisfaction values calculated for link costs corresponding to present and future scenarios are used to evaluate the users' benefits (Chapter 10).

Section 7.4 describes some simple algorithms for stochastic and deterministic equilibrium assignment models with rigid demand. Section 7.5 describes algorithms for the determination of the shortest hyperpaths and their application to traffic assignment with pre-trip/en-route path choice models. Finally, section 7.6 describes the extension of the algorithms presented in the previous sections to multi-mode

(°) Giulio Erberto Cantarella co-author of this chapter.

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

assignment with elastic demand. All presented algorithms can easily be extended for multi-user assignment, which will not be explicitly addressed.

In what follows, it is assumed that the network topology is known and that each origin/destination pair is connected . In some cases, apart from the notation adopted in Chapter 2, it will be useful to indicate explicitly the end nodes of a link $l \equiv (i,j)$.

7.2. Shortest path algorithms

The identification of the shortest paths between pairs of centroids is used in the simulation of path choice behavior within the assignment algorithms. In particular, deterministic path choice simulation requires the identification of the shortest path (or paths) between each pair of centroids, while in probabilistic path choice the shortest paths are sometimes inputs of stochastic uncongested network assignment algorithms. Furthermore, most models used for the construction of the set of relevant paths, with a selective approach and explicit paths enumeration (described in section 4.2.4.1), lead to the solution of a shortest path problem. (Such a set can be specified shortest path algorithms with respect to different link attributes such as distance, monetary cost, travel time, or through the identification of the first k shortest paths).

Assuming that only *elementary paths*, i.e. those without loops, are relevant, there is a finite number of such paths and they could be enumerated for each O-D pair. When the explicit enumeration of all paths is not feasible due to the large number, as it is often the case, shortest path algorithms that avoid explicit path enumeration have to be adopted as described in this section.

Applications relative to transportation network assignment do not require the determination of the shortest path between all the possible pairs of nodes, only between centroids. This section, therefore, describes the basic structure of shortest paths algorithms from an origin centroid o to all the network nodes (*forward shortest paths*) or from all the network nodes to a destination centroid d (*backward shortest paths*)[(1)].

In what follows it is assumed that shortest paths cannot cross centroids; as such each centroid can be represented in the network model by two unconnected nodes, the origin, from which links exit, and the destination, into which links enter[(2)]. Furthermore, for the sake of simplicity, the variable, assumed non-negative[(3)], assigned to each link will be called cost, since it usually represents the generalized transportation cost; however it could be any other performance variable (distance, travel time etc.). Let

$c_l = c_{ij} \geq 0$ be the cost on link $l = (i,j)$;

$Z_{i,j} \geq 0$ be the cost of the shortest path between the any pair of nodes i and j; note that in general it may result $Z_{i,j} \neq Z_{i,j}$ in (e.g. due to one-way streets, different slopes, etc.).

Shortest path costs satisfy the *triangular inequality*:

$$Z_{o,i} + Z_{i,d} \geq Z_{o,d} \qquad \forall i \ \forall od$$

In fact, if for an od pair there was a node i for which $Z_{o,i} + Z_{i,d} < Z_{o,d}$, the cost of the path from o to d through the node i would be less than $Z_{o,d}$, against the assumption that $Z_{o,d}$ is the cost of the shortest path from o to d.

The triangular inequality implies that link costs and shortest path costs satisfy the *Bellman principle*' which states that *a shortest path is made up by shortest paths*:

if the link *(i,j)* belongs to the shortest path between o and j:
then $Z_{o,i} + c_{ij} = Z_{o,j}$ otherwise $Z_{o,i} + c_{ij} \geq Z_{o,j}$

More in general:

if the link *(i,j)* belongs to the shortest path between o and d
then $Z_{o,i} + c_{ij} + Z_{j,d} = Z_{o,d}$ otherwise $Z_{o,i} + c_{ij} + Z_{j,d} \geq Z_{o,d}$

It can easily be seen that the Bellmann principle is equivalent to the first Wardrop principle (for a uncongested network), discussed in Section 5.3.2. If there is only one shortest path between each pair of nodes, the second condition of each of the two formulations of the Bellman principle holds as strict inequalities.

It should be recognized that if there is only one shortest path between each pair of nodes (or if there are several shortest paths, only one is taken into consideration), the shortest paths from an origin node o to other nodes form a *forward tree*[4], $T(o)$, with root o. Any forward tree $T(o)$ can be described by the link, necessarily unique, entering each node j (or from the initial node of this link). Similarly, the shortest paths from all the nodes to a destination node o form a backward tree, $T(d)$, with root d. Any tree $T(d)$ can be described by the link, necessarily unique, exiting from each node i (or from the final node of this link). (The use of the same notation for forward trees from an origin o and backward trees towards a destination d does not generate ambiguity if only trees relative to centroids are considered. In this case the type of node n, origin or destination, defines the type of tree, forward and backward.)

Given a forward tree $T(o)$ from the origin node o, let

$X_{T(o),i} \geq 0$ be the cost along the only path from the origin node o to the node i in the tree $T(o)$.

It yields: $$X_{T(o),i} + c_{ij} = X_{T(o),j} \qquad \forall (i,j) \in T(o)$$

The tree $T(o)$ from the origin node o is the tree of shortest paths (or one of such trees if there are several shortest paths between some pairs of nodes) if and only if the following condition deduced from the Bellman principle, holds (see example in Fig. 7.2.1a):

$$X_{T(o),i} + c_{ij} \geq X_{T(o),j} \qquad \forall (i,j) \notin T(o) \qquad (7.2.1)$$

In this case, the values $X_{T(o),i}$ are the shortest path costs $Z_{o,i}$.

Similarly, given a backward tree $T(d)$ from the destination node d, let

$X_{i,T(d)} \geq 0$ be the cost along the unique path from the node i to the destination d in the tree $T(d)$.

It yields: $$c_{ij} + X_{j,T(d)} = X_{i,T(d)} \qquad \forall (i,j) \in T(d)$$

The tree $T(d)$ to the destination node d is the tree of shortest paths (or one of such trees if there are several shortest paths between some pairs of nodes) if and only if the following condition deduced from the Bellman principle, holds (see example in Fig. 7.2.1b):

$$c_{ij} + X_{j,T(d)} \geq X_{i,T(d)} \qquad \forall (i,j) \notin T(d) \tag{7.2.1}$$

In this case the values $X_{i,T(d)}$ are the shortest path costs $Z_{i,d}$.

The algorithms used to identify the shortest paths are based on the iterative updating of tentative shortest path tree, until condition (7.2.1) or (7.2.2) hold. Examples of forward update for the forward tree from the origin o, and for the backward tree towards the destination d, are shown in Fig. 7.2.2a and Fig. 7.2.2b respectively.

The algorithms based on the described updating steps stop when not further updates can be performed. The number of steps depends on the node-list management strategy i.e. the strategy for the choice of the node used to verify whether further updating steps are needed. Note that, under the assumption of non-negative costs, given a tentative tree $T(o)$ from the origin o, the cost $X_{T(o),i}$ along the unique path from the origin o to the node i cannot undergo further updating starting from another node j with a higher guess value, $X_{T(o),j} > X_{T(o),i}$. Similarly, for a tentative tree $T(d)$ towards the destination d.

Using the above considerations, algorithms with ordering (know also in literature as Label-Setting) at each iteration make definitive the node i with the lowest value $X_{T(o),i}$ among those which are not yet definitive. Then updating step are performed from this node . The algorithm requires as many updating steps as there are nodes, since each step makes the tentative value of a node definitive. Note also that the nodes are made definitive in order of increasing shortest path cost. These algorithms require carefully designed data structures to handle the ordering of nodes to be examined. On the other hand, algorithms without ordering (know also in literature as Label-Correcting) generally are simpler to implement, but all tentative values become definitive only at the end of the algorithm. In this case the (finite) number of updating steps depends on the list management strategy adopted.

When there are multiple shortest paths, the algorithms described yield a unique path depending on the order in which the nodes are examined. The algorithms can be modified to give all the possible shortest paths. In this case, however, the set of shortest paths from an origin (or towards a destination) is no longer a tree.

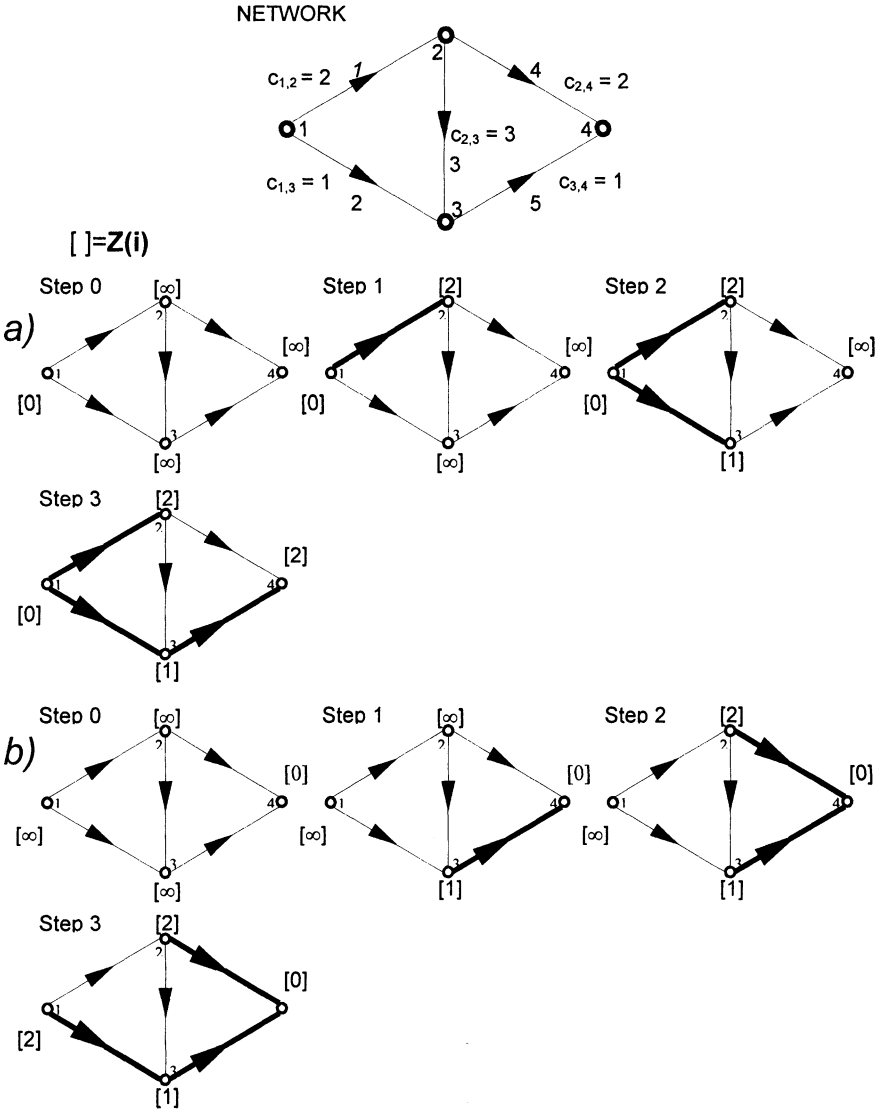


Fig. 7.2.1 Example of forward (a) and backward (b) shortest path algorithms.

7.3. Algorithms for Uncongested Network assignment

Uncongested Network (UN) assignment models simulate systems with link and path costs independent of flows. As stated in Chapter 5, this type of assignment is adopted for the analysis of moderality congested road and public transportation systems. Uncongested network assignment algorithms are also used within equilibrium assignment algorithms described in the following sections.

In the case of *explicit path enumeration*, path costs can be easily computed from link costs through the link-path incidence relationship (5.2.1). Similarly path flows can be obtained by applying the demand model (5.2.7) and its extensions, and link flows can be computed from path flows using the congruence relationship (5.2.3). It is also possible to calculate the EMPU variable, $s_{od} = s_{od}(-\Delta_{od}^T \boldsymbol{c} - \boldsymbol{g}_{od}^{NA})$, related to path choice for each *od* pair.

However, it should be recalled that for Probit path choice models, it is not possible to calculate analytically choice probabilities and therefore the demand model (5.2.7) and its extensions. Unbiased estimates of path choice probabilities, and of the corresponding path flows, can be obtained by using a Monte Carlo sampling technique[5] of paths random residuals. Given a sample of perceived path cost vectors, for each path cost vector the demand flow of each O-D pair is assigned to the (perceived) shortest path. The average of path flows obtained for the different cost vectors in the sample is an unbiased estimate of the Stochastic Uncongested Network path flows. It yields:

$$\bar{\boldsymbol{h}}^m = \Sigma_{j=1,m} \boldsymbol{h}^j / m$$

where

$\boldsymbol{h}^j = \boldsymbol{h}_{SPA}(\boldsymbol{g} + \boldsymbol{\varepsilon}^j)$ is the vector of the path flows obtained by assigning the demand flow of each O-D pair to the shortest path with respect to path costs $\boldsymbol{g} + \boldsymbol{\varepsilon}^j$;

$\boldsymbol{g}$ is the vector of (systematic) costs vector of the paths;

$\boldsymbol{\varepsilon}^j \leftarrow MVN(\boldsymbol{0},\Sigma)$ is the j-th vector of random residuals in a sample of m vectors, $\boldsymbol{\varepsilon}^j$ is obtained as pseudo-realization of a normal multivariate random variable with null mean and variance-covariance matrix Σ;

$\bar{\boldsymbol{h}}^m$ is an unbiased estimate of the SUN assignment path flows vector, obtained with a sample of m vectors of perceived path costs.

From a practical point of view, the path flow estimate $\bar{\boldsymbol{h}}^m$ can be obtained through the following recursive equations up to $j = m$, assuming initially $j = 0$ and $\bar{\boldsymbol{h}}^0 = 0$:

$$j = j+1$$
$$\boldsymbol{\varepsilon}^j \leftarrow MVN(\boldsymbol{0},\Sigma)$$
$$\boldsymbol{h}^j = \boldsymbol{h}_{SPA}(\boldsymbol{g} + \boldsymbol{\varepsilon}^j)$$
$$\bar{\boldsymbol{h}}^j = ((j\text{-}1)\ \bar{\boldsymbol{h}}^{j-1} + \boldsymbol{h}^j) / j$$

For each *od* pair, the average perceived shortest path costs is an unbiased estimate of the EMPU variable associated to path choice. Direct use of this approach is complex in applications because of the need to generate realizations of a Multivariate Normal with non-null covariances, $\varepsilon^j \leftarrow MVN(\mathbf{0},\Sigma)$. It is therefore convenient to generate perceived path costs from link costs, adopting the same approach described in Section 6.3.2b in the case of implicit paths enumeration.

It is also possible to calculate *UN* assignment link flows *without explicit path enumeration,* in the absence of non-additive path costs, using procedures based on shortest path algorithms as described below, first for *SUN* and then for *DUN* assignment.

7.4.1. Stochastic Uncongested Network assignment without explicit paths enumeration

In Stochastic Uncongested Network (SUN) assignment it is assumed that each user associates to each path connecting the O-D pair a perceived utility represented by a random variable whose expected value is given by minus the path cost (see Section 4.5.2)[6]. Different SUN assignment models can be specified according to the specification of path choice model. In the following an algorithm for Logit SUN assignment without explicit paths enumeration is described first. Successively an algorithm for Probit SUN assignment without explicit paths enumeration is reported.

7.3.1.1. SUN assignment with Logit path choice model

For Logit path choice models, it is possible to compute link flows without explicit paths enumeration with an algorithm known in the literature as *Dial algorithm* after its author. This algorithm is associated to a particular specification of the Logit path choice model, where only *efficient paths with respect to the origins*, belong to the set of relevant paths; these paths are made up by links *l=(i,j)* termed *efficient links*, such that the cost of the shortest path to reach the initial node *i* from the origin *o* is smaller than the cost of the shortest path to reach the final node *j*, say $Z_{o,i} < Z_{o,j}$ Assuming that the link costs are strictly positive, $c_{ij} > 0$, the links of the shortest paths tree are efficient by definition and, therefore, the shortest paths are among the efficient paths. Thus the efficiency condition must be tested only for links not belonging to the shortest path tree. Analogously, *efficient paths with respect to the destinations* can be defined. It is also possible to define *efficient paths with respect to both the origin and the destination*; in this case, each O-D pair must be analyzed separately, with a lower computational efficiency. This definition of the set of relevant paths implies that a selective approach is adopted for choice set definition, according to the Dial efficiency criterion, classified as topological in section 4.3.4. For brevity, only the case of efficient paths with respect to the origins will be described.

Fig. 7.3.1 illustrates some efficient paths from origin 1 to destination 4. Notice that with the first configuration only the shortest paths are efficient. This is no

longer the case for costs b) and c). The examples show that efficiency does not depend on topology only.

The theoretical analysis of the Dial algorithm is based on an equivalent formulation of the Logit model, highlighting the role of link costs in the specification of path costs. This formulation allows simultaneous analysis of all the O-D pairs with a common origin o. In particular, as was seen in Section 4.3.4.1 the Logit probability $p_{od,k}$ of choosing path k for users traveling from origin o to destination d is given by:

$$p_{od,k} = exp(-g_k/\theta) / \Sigma_{j \in Iod} exp(-g_h/\theta) \propto exp(-g_k/\theta) \qquad (7.3.1)$$

where

$\theta = (\sqrt{6}/\pi)\sigma$ is the parameter of the Logit model, proportional to the standard deviation of random residuals;
g_k is the cost on the path k;
K_{od} is the set of the (relevant) paths connecting the pair od.

If (additive) path costs are expressed as the sum of link costs through the congruence relationship (5.2.1), expression (7.3.1) yield:

$$p_{od,k} \propto exp(-\Sigma_{(i,j)\in k} c_{ij}/\theta) = \Pi_{(i,j)\in k} exp(-c_{ij}/\theta) \qquad (7.3.2)$$

More generally, if each path is considered to be a sequence of nodes, j, and of links (i,j), the probability $p_{od,k}$ can be expressed as the product of the probabilities, $Pr[(i,j)/j]$, of choosing each link (i,j) of the path k conditional on the final node j (Fig. 7.3.2):

$$p_{od,k} = \Pi_{(i,j)\in k} Pr[(i,j)/j] \qquad (7.3.3)$$

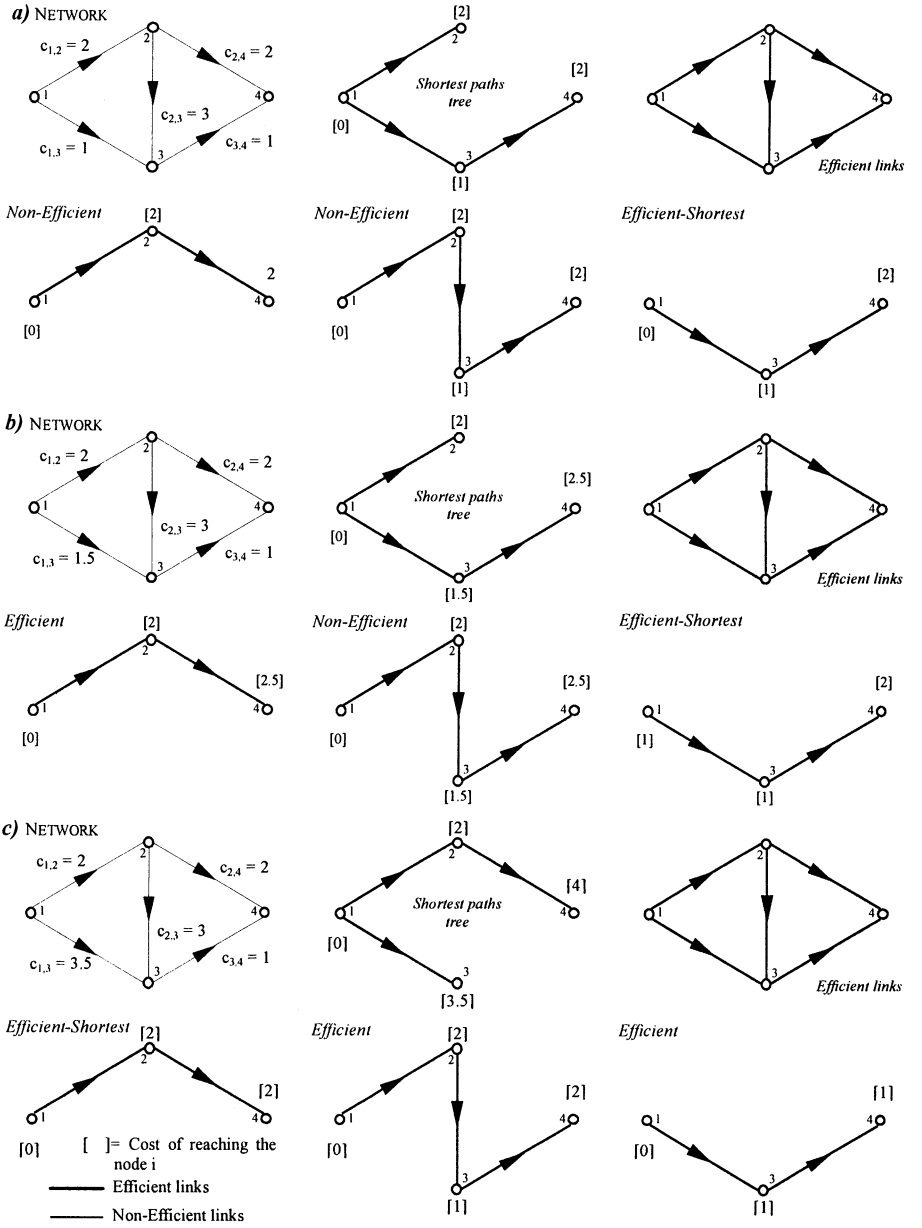


Fig. 7.3.1 Examples of efficient paths.

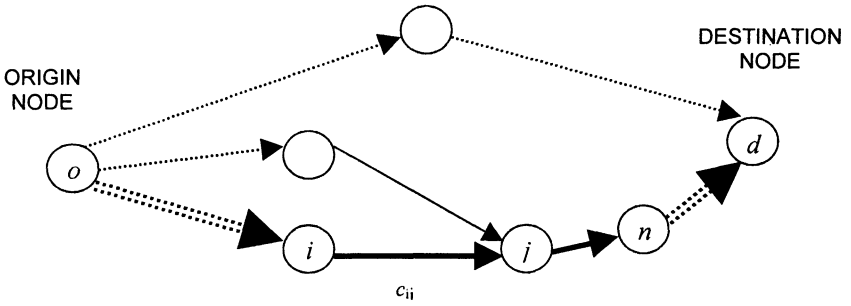


Fig. 7.3.2 Path k from the origin o to the destination d through the link (i,j).

The probability $p_{od,k}$ calculated with (7.3.3) coincides with (7.3.2) if the probability $Pr[(i,j) / j]$, of choosing the link (i,j) conditional on the final node j is defined with a Logit model of parameter θ. In this model, the alternatives are the efficient links (i,j) incident to node j (i.e. all the efficient links entering node j) and the systematic utility of each alternative $V_{ij/j}$ is the sum of the opposite of the link cost c_{ij} and of a logsum variable Y_i synthetically taking into account the utilities of all the efficient paths from the origin o to the initial node i of the link:

$$Pr[(i,j) / j] = exp(V_{ij/j}/\theta) / \Sigma_{(m,j)\in BS(j)} exp(V_{mj/j}/\theta) \quad (7.3.4)$$

$$V_{ij/j} = -c_{ij} + \theta Y_i \quad (7.3.5)$$

$$Y_i = ln(\Sigma_{(n,i)\in BS(i)} exp(V_{ni/i}/\theta)) \quad (7.3.6)$$

where
$BS(j)$ is the backward star of node j, i.e. the set of links (i,j) incident in the node j;
Y_i is the logsum variable of the utilities of the incident links in the node i.

The relationships (7.3.4-6) yields:

$$Pr[(i,j) / j] = exp((-c_{ij} + \theta Y_i)/\theta) / \Sigma_{(m,j)\in BS(j)} exp((-c_{mj} + \theta Y_m) /\theta) = w_{ij} / W_j$$

with

$$w_{ij} = exp((-c_{ij} + \theta Y_i)/\theta) = exp(-c_{ij}/\theta)\, exp(Y_i) = exp(-c_{ij}/\theta)\Sigma_{(n,i)\in BS(i)} \exp(V_{ni/i}/\theta)$$

$$W_j = \Sigma_{(m,j)\in BS(j)} exp((-c_{mj} + \theta Y_m)/\theta) = \Sigma_{(m,j)\in BS(j)} w_{ij}$$

The probability $Pr[(i,j) / j]$ of choosing the link (i,j) conditional on the final node j can therefore be expressed as the ratio between a weight w_{ij} associated to the link (i,j), and a weight W_j associated to the node j. Note that the definition of the link weights yields:

$$w_{ij} = exp(-c_{ij}/\theta)\Sigma_{(n,i)\in BS(i)} exp(V_{ni/i}/\theta) = exp(-c_{ij}/\theta)W_i$$

Furthermore, a null weight, $w_{ij} = 0$, is associated to non-efficient links i.e. the links (i,j), with $Z_{o,i} \geq Z_{o,j}$, consistently with the assumption that a link (i,j) belongs to

a path if and only if the shortest path to reach its initial node from the origin is less than the shortest path to reach its final node.

From the above considerations, the weights of the links, w_{ij}, of the nodes, W_j, and the probabilities $Pr[(i,j)/j]$ can be determined by using recursive equations equivalent to the relations (7.3.4-6). They are computed for each link, starting from the origin o, with $W_o = 1$, and continuing with the other nodes i by increasing minimum cost $Z_{o,i}$:

$$w_{ij} = \begin{cases} exp(-c_{ij}/\theta)\, W_i & \text{if } Z_{o,i} < Z_{o,j} \\ 0 & \text{if } Z_{o,i} \geq Z_{o,j} \end{cases} \tag{7.3.7}$$

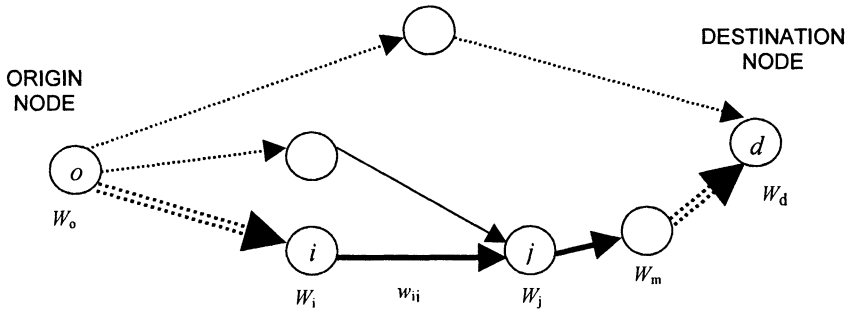
$$W_j = \Sigma_{(m,j)\in BS(j)}\ w_{mj} \tag{7.3.8}$$

$$Pr[(i,j)/j] = w_{ij} / W_j \tag{7.3.9}$$

Substituting the relationships (7.3.7-9) in (7.3.3) yields expression (7.3.2).

In fact, the weights of the path nodes, excluding the origin and the destination, are irrelevant since they appear both in the numerator and in the denominator, as the final node of a link is the initial node of the next link along the path (see Fig. 7.3.3):

$$p_{od,k} = \Pi_{(i,j)\in k}\, W_i\, exp(-c_{ij}/\theta)/W_j = \\ = \Pi_{(i,j)\in k}\, exp(-c_{ij}/\theta)\, W_o/W_d \propto \Pi_{(i,j)\in k}\, exp(-c_{ij}/\theta)$$



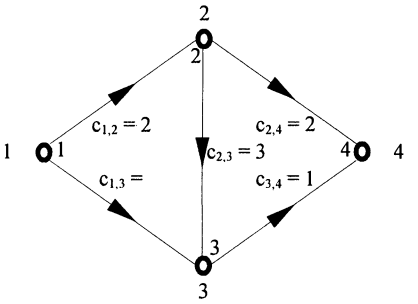
$$p_{od,k} = \frac{W_o\, exp(-\alpha c_{oa})}{W_a} \cdots \times \cdots \frac{W_i\, exp(-\alpha c_{ij})}{W_j} \times \frac{W_j\, exp(-\alpha c_{jm})}{W_m} \cdots \times \cdots \frac{W_z\, exp(-\alpha c_{zd})}{W_i}$$

Fig. 7.3.3 Node and link weights.

The Dial algorithm for SUN assignment is based on the iterative calculation of the weights of the nodes and links, for each origin o, using the relationships (7.3.7) and (7.3.8). The processing of nodes by increasing minimum cost ensures that it is possible to apply the recursive relationships (7.3.7) and (7.3.8), i.e. that when the weight w_{ij} of a link (i,j), is computed the weight W_i of the initial node i has already been determined. When the weights of all the nodes and links are known, the demand flow d_{od} from each destination d is assigned backward to the various links with link probabilities given by expression (7.3.9). Given an origin o, for each destination d, the EMPU value for path choice is given by the inclusive variable relative to the destination d, $s_{od} = Y_d$. An example is given in Fig. 7.3.4. The calculation time is two or three times greater than the time needed for DUN assignment.

The algorithm described can be extended to calculate SUN assignment link flows for C-Logit path choice models (described in Section 4.2.5.1) given that one O-D pair is examined at a time and an appropriate specification of the commonality factor is adopted.

Observe that the shortest paths used to define efficient paths can be calculated with a vector of link cost attributes different from link costs c. For example, it can be assumed that efficient paths are defined in terms of their physical length (or another attribute), while a cost proportional to the travel time is used to simulate users' choice among these paths. In this case the shortest paths and the distances $Z_{o,i}$ are calculated with the physical lengths of the links, while link weights w_{ij} and node weights W_i are calculated using the costs (times)$^{(7)}$ c_{ij}. Under this assumption the set of efficient paths is independent of link costs. This feature is particularly relevant for stochastic equilibrium assignment since for congested networks the SUN function is increasing monotone in terms of the (congested) link costs and has a symmetric Jacobian (see section 5.3.1). Therefore, the (sufficient) condition for stochastic equilibrium uniqueness is ensured, as is the convergence of stochastic equilibrium algorithms described in Section 7.4.2.

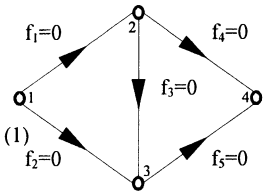


$1 \rightarrow 4$: $d_{14}=1000$

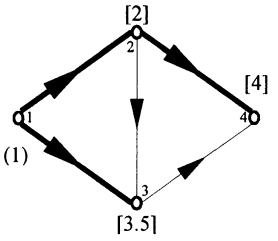
Cost	Ord(k)
[0]	1
[2]	2
[3.5]	3
[4]	4

$\theta=2$

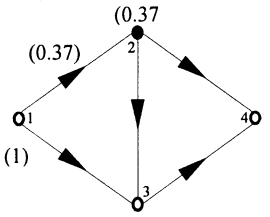
INITIALIZATION



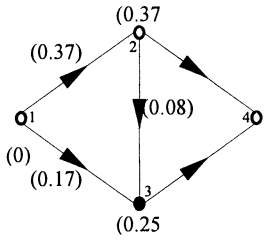
SHORTEST PATHS TREE



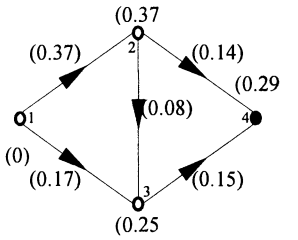
WEIGHTS AT NODE 2



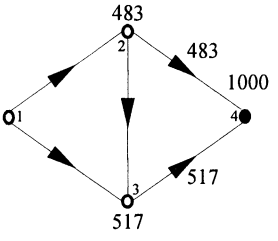
WEIGHTS AT NODE 3



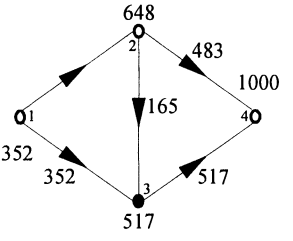
WEIGHTS AT NODE 4



FLOWS AT NODE 4



FLOWS AT NODE 3



FLOWS AT NODE 2

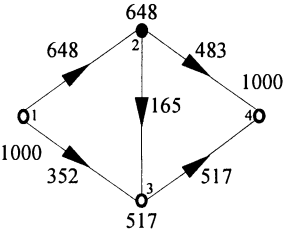


Fig. 7.3.4 Application of the Dial algorithm for Logit SUN assignment.

7.3.1.2. SUN assignment with Probit path choice model

The Probit path choice model results from the assumption that the random residuals, elements of the vector ε, are distributed according to a Multivariate Normal, $MVN(\mathbf{0}, \Sigma)$, with null mean and variance-covariance matrix Σ. This model can account for overlapping paths, introducing a positive covariance between the perceived utilities of two paths sharing some links, but it does not allow explicit calculation of paths choice probabilities. Unbiased estimates of path choice probabilities and their corresponding SUN path and link flows can be obtained by using a Monte Carlo technique.

An algorithm not requiring explicit path enumeration can be specified with a particular specification of the path choice model, assuming that each user associates to each path a perceived utility represented by a random variable with expected value given by the opposite of path cost (see section 4.2.5.1):

$$\boldsymbol{U} = \boldsymbol{V} + \boldsymbol{\varepsilon} = -\boldsymbol{g} + \boldsymbol{\varepsilon} \tag{7.3.10}$$

with

$$E[\boldsymbol{g}] = \boldsymbol{g} = -\boldsymbol{V} = -E[\boldsymbol{U}] \qquad Var[\boldsymbol{g}] = \mathbf{0}$$
$$E[\boldsymbol{\varepsilon}] = \mathbf{0} \qquad Var[\boldsymbol{\varepsilon}] = Var[\boldsymbol{U}] = \boldsymbol{\Sigma}$$

where

$\boldsymbol{U}$ is the vector of the perceived path utility, with expected value $\boldsymbol{V} = E[\boldsymbol{U}]$, and variance-covariance matrix $Var[\boldsymbol{U}] = \boldsymbol{\Sigma}$;

$\boldsymbol{g} = -\boldsymbol{V}$ is the path costs vector, given by minus the systematic utility vector, $\boldsymbol{V}$;

$\boldsymbol{\varepsilon} = \boldsymbol{U} - E[\boldsymbol{U}]$ is the vector of random residuals relative to the path utilities.

The congruence between link and path costs through the link-path incidence matrix, $\boldsymbol{\Delta}$, and the assumption of purely additive path costs allow to express the relationship (7.3.10) in terms of link utility, costs and random residuals. Let

$\boldsymbol{U}$ be the vector of link perceived utilities, with expected value $\boldsymbol{v} = E[\boldsymbol{u}]$ and variance-covariance matrix $Var[\boldsymbol{u}] = \boldsymbol{\Sigma}_l$;

$\boldsymbol{c} = -\boldsymbol{v}$ be the vector of the link costs, assumed to be equal to minus the link systematic utility $\boldsymbol{c} = -\boldsymbol{v}$;

$\boldsymbol{\eta} = \boldsymbol{u} - E[\boldsymbol{u}]$ be the vector of random residuals relative to link utilities.

It results:

$$\boldsymbol{U} = \boldsymbol{\Delta}^T \boldsymbol{u} \tag{7.3.11}$$
$$\boldsymbol{g} = \boldsymbol{\Delta}^T \boldsymbol{c} \tag{7.3.12}$$
$$\boldsymbol{\varepsilon} = \boldsymbol{\Delta}^T \boldsymbol{\eta} \tag{7.3.13}$$

thus

$$\boldsymbol{u} = -\boldsymbol{c} + \boldsymbol{\eta} \tag{7.3.14}$$

with

$$E[\boldsymbol{c}] = \boldsymbol{c} = -E[\boldsymbol{u}] \qquad Var[\boldsymbol{c}] = \mathbf{0}$$
$$E[\boldsymbol{\eta}] = \mathbf{0} \qquad Var[\boldsymbol{\eta}] = Var[\boldsymbol{u}] = \Sigma_l$$

Since the relationships (7.3.11 – 13) are linear, the variance-covariance matrix of paths random residuals, Σ, depends on the variance-covariance matrix of link random residuals, Σ_l, through the relationship:

$$\Sigma = \Delta^T \Sigma_l \Delta \tag{7.3.15}$$

These results can be interpreted as a specification of the path choice model in which users perceive the costs of individual links and the perceived cost of a path is given by the sum of perceived link costs.

By using the above approach, an algorithm can be specified that does not require explicit path enumerationbased on the assumption that the choice set consists of all the elementary paths, and that the variance-covariance matrix Σ has the structure described in Section 4.2.5.1. Let

g_k	be the cost of path k;
g_{kj}	be the cost on the links shared by the paths k and j;
$\sigma_{kk} = \sigma_k^2$	be the variance of the random residual of path k, an element of the main diagonal of the variance-covariance matrix Σ;
σ_{kj}	be the covariance between the random residuals of paths k and j, element of the variance-covariance matrix Σ;
ξ	be the proportionality coefficient between path costs and elements of the variance-covariance matrix (expressed in units coherent with costs and utilities).

Under the quoted assumptions made on the structure of the variance-covariance matrix, it yields:

$$\sigma_k^2 = \sigma_{kk} = \xi\, g_k$$
$$\sigma_{kj} = \xi\, g_{kj}$$

Consistent with the relationship of congruence between link and path costs, it results:

$$g_k = \Sigma_l\, \delta_{lk}\, c_l = \Sigma_l\, \delta_{lk}^2\, c_l$$
$$g_{jk} = \Sigma_l\, \delta_{lk}\, a_{lj}\, c_l$$

thus

$$\sigma_{kk} = \xi\, \Sigma_l\, \delta_{lk}^2\, c_l$$
$$\sigma_{kj} = \xi\, \Sigma_l\, \delta_{lk}\, \delta_{lj}\, c_l$$

Indicated by $\boldsymbol{DIAG}(c)$ the diagonal matrix with elements on the main diagonal given by link costs, $\boldsymbol{c}$, in matrix terms it results:

$$\Sigma = \xi \Delta^T \boldsymbol{DIAG}(c)\Delta \tag{7.3.16}$$

To achieve the above condition it may be assumed that each link random residual, η_l, is independently distributed according to a univariate Normal $N(0, \sigma_l^2)$ with null mean and variance $\sigma_l^2 = \xi c_l$. Therefore the vector $\boldsymbol{\eta}$ is distributed according to a multivariate Normal $MVN(\mathbf{0}, \Sigma_l)$ with null mean and diagonal variance-covariance matrix defined by:

$$\Sigma_l = \xi \boldsymbol{DIAG}(c) \tag{7.3.17}$$

In this case the path random residuals deriving from the linear relationship (7.3.13), $\boldsymbol{\varepsilon} = \Delta^T \boldsymbol{\eta}$, are distributed according to a multivariate Normal, $MVN(\mathbf{0}, \Sigma)$, with variance-covariance matrix given by the relationship (7.3.15), which combined with (7.3.17), provides the relationship (7.3.16).

Therefore, a sample of vectors of normally distributed path random residuals, $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \Sigma)$, can be obtained starting from a sample of vectors of links random residuals $\boldsymbol{\eta}$, resulting from independently sampling the residual, η_l, of each link l from a univariate normal, $N(0, \sigma_l^2)$.

It should be stressed that the link costs used to define the variance-covariance matrix through the relation (7.3.16) may be different from the actual link costs $\boldsymbol{c}$, expressing the systematic utility of the link, $\boldsymbol{v} = -\boldsymbol{c}$, and therefore of the path, $\boldsymbol{g} = -\boldsymbol{V}$. For example it can be assumed that the similar perception of two overlapping paths, expressed by the covariance of their random residuals, is proportional to the length of the links shared, while the systematic link cost is a function of the travel time (dependent on flows for congested networks). These assumptions ensure that the SUN assignment function is non-increasing monotone with respect to (congested) link costs and has symmetric Jacobian (section 5.3.1). The (sufficient) condition for the resulting stochastic equilibrium uniqueness is therefore ensured (as described in section 5.4.1), as is the convergence of stochastic equilibrium algorithms described in section 7.4.2.

From an algorithmic point of view, in order to calculate SUN assignment flows with a Probit path choice model, a sample of perceived link cost vectors has to be generated. For each sample of (perceived) link costs, demand flows are assigned to the (perceived) shortest paths with a DUN (All or Nothing) assignment algorithm, described in the next subsection. The mean of the link flows obtained for the different link cost vectors of the sample is an unbiased estimate of Probit SUN link flows. The algorithm can be stated formally by introducing the following variables:

$\eta_l^j \leftarrow N(0, \sigma_l^2 = \xi\, c_l)$ the j-th random residual for link l, in a sample of m, obtained as a pseudo-realization of a normal random variable with zero mean and variance $\sigma_l^2 = \xi\, c_l$;

$r_l^j = c_l + \eta_l^j$ the j-th perceived cost for the link l, in a sample of m;

$\boldsymbol{r}^j = [r_l^j]_l$ the j-th vector of perceived link costs, with elements r_l^j;

$\boldsymbol{f}^j = \boldsymbol{f}_{DUN}(\boldsymbol{r}^j)$ the Deterministic Uncongested Network assignment link flow vector corresponding to link costs $\boldsymbol{r}^j$ (computed as described in next subsection);

$\bar{\boldsymbol{f}}^m$ an unbiased estimate of the vector of Stochastic Uncongested Network assignment link flows, obtained with a sample of m vectors of perceived link costs.

It yields:

$$\bar{\boldsymbol{f}}^m = \Sigma_{j=1,m} \boldsymbol{f}^j \;/\; m$$

From a practical point of view, the link flows estimate, $\bar{\boldsymbol{f}}^m$, can be obtained with the following recursive equations up to $j = m$, initially supposing $j = 0$ and , $\bar{\boldsymbol{f}}^0 = 0$:

$$\begin{aligned} &j = j+1 \\ &\eta_l^j \leftarrow N(0, \sigma_l^2 = \xi\, c_l) \;\; \forall l \\ &\boldsymbol{r}^j = [c_l + \eta_l^j]_l \\ &\bar{\boldsymbol{f}}^m = ((j\text{-}1)\, \bar{\boldsymbol{f}}^m + \boldsymbol{f}_{DUN}(\boldsymbol{r}^j)) / j \end{aligned}$$

For each pair od, the average of the minimum costs obtained with the different shortest paths, Z_{od}^j, is an unbiased estimate of the opposite of the EMPU variable for path choice $\bar{s}_{od}^{\,m} = -\Sigma_{j=1,m} Z_{od}^j \;/\; m$.

This algorithm, also known in the literature as *Monte Carlo* , unlike other algorithms in this chapter, does not yield link flow values, but only a sequence of unbiased estimates whose precision increases with the number of iterations.

In practice, the algorithm continues until a stop criterion is met e.g. a pre-assigned maximum number of iterations, j_{max}. The algorithm could also terminate when the difference between the link flows estimates in two successive iterations is below a pre-assigned threshold δ, by using a suitable norm $| \bar{\boldsymbol{f}}^j - \bar{\boldsymbol{f}}^{j-1}| / | \bar{\boldsymbol{f}}^{j-1}| < \delta$, or on a link basis $| \bar{f}_l^j - \bar{f}_l^{j-1}| / | \bar{f}_l^{j-1}| < \delta$. However, this criterion is not very effective since as the number of iterations j increases, it tends to be verified in any case, and it is in practice the same as supposing a maximum number of iterations. More correctly the algorithm should be stopped when the sample estimate of the precision of link flows is below a given threshold, $\max_l [\ \mathrm{var}(\bar{f}_l^{\,m})^{(1/2)} / \bar{f}_l^{\,m}\ ] \leq \delta$. Alternatively, a statistical equality test between two successive averages can be used. It can easily be proved that whatever the convergence criterion adopted, the calculation time is roughly equal to m times the time needed to carry out a Deterministic Uncongested Network assignment (with any of the algorithms described in the next subsection).

An example of Monte Carlo algorithm is given in Fig. 7.3.6.

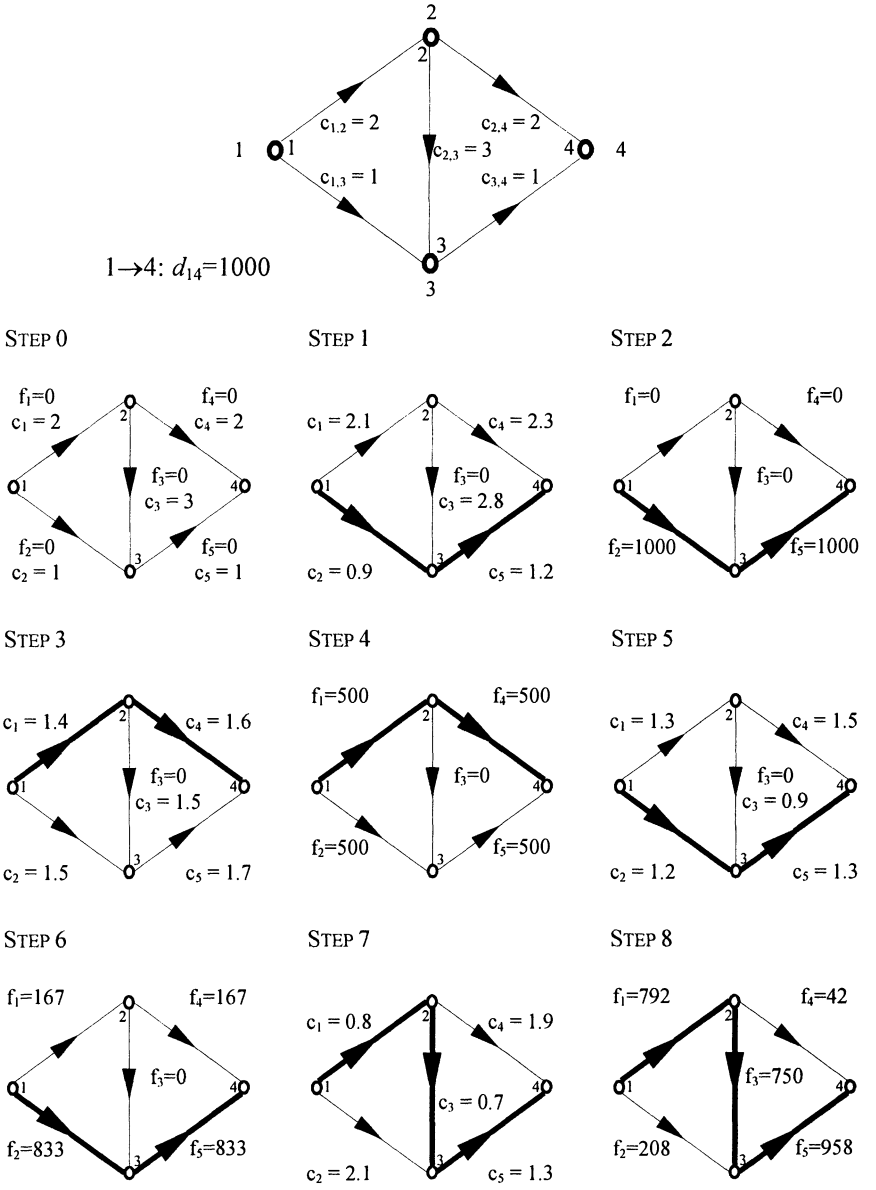


Fig. 7.3.6 Example of the Monte Carlo algorithm for Probit SUN assignment

7.3.2. Deterministic Uncongested Network assignment without explicit paths enumeration

Under the assumption of deterministic path choice behavior, all users traveling between each O-D pair choose the shortest path (Section 5.3.2) leading to Deterministic Uncongested Network (DUN) assignment. As observed, if there are several shortest paths for each O-D pair, path flows, and therefore link flows, are not uniquely defined. However, shortest path algorithms provide a unique path between each O-D pair. This path depends on the implementation details of the algorithm and in particular on the ordering of the nodes. Link flows can therefore be calculated by assigning the entire demand flow of each O-D pair to the links of the shortest path generated by the algorithm and zero to the links of all the other paths. These algorithms are known as *All-or-Nothing* and can be implemented following two different approaches.

In the *sequential* approach, once the shortest path tree from an origin o has been calculated, the demand d_{od} towards each destination d is added to the flows on all the path links from o to d. An example of sequential algorithm is given in Fig. 7.3.8. The procedure is analogous if the shortest path tree towards each destination d is calculated.

Instead of the simple algorithms described above, other algorithms following a *simultaneous* approach can be used. Simultaneous algorithms are computationally more efficient and can be extended to DUN assignment models for transit networks (shortest hyperpaths) as described in section 7.5. These algorithms are particularly efficient if, for each shortest path tree, there is a list of nodes ordered by increasing minimum cost from the origin (or to the destination). Such an order can easily be obtained by applying shortest path algorithms with ordering.

Simultaneous algorithms from an origin are based on the calculation of the flow entering each node, defined as the sum of the flows on the links entering the node. Considering one origin o at a time, the demand flow d_{od} is initially assigned as the flow entering each destination and a zero tentative flow is assigned as entering flow in all other nodes. Once the tree of the shortest paths from the origin o has been calculated, each node i in decreasing minimum cost order is examined, starting from the furthest node from the origin o (i.e. the node with to the highest value Z_{oi}) until the origin o is reached. The flow entering each node i is assigned to the unique previous link in the tree, and added to the flow entering the initial node of this link. The order adopted is such that when a node i is examined all the furthest nodes have already been examined. Therefore there cannot be any node still to be examined from which the flow contributes to the flow entering the node i[8]. For each od pair, the EMPU related to deterministic path choice is given by the cost on the shortest path, $s_{od} = Z_{od}$.

An example of the application of a simultaneous algorithm is given in Fig. 7.3.11. The procedure is analogous if shortest path trees towards each destination d are calculated.

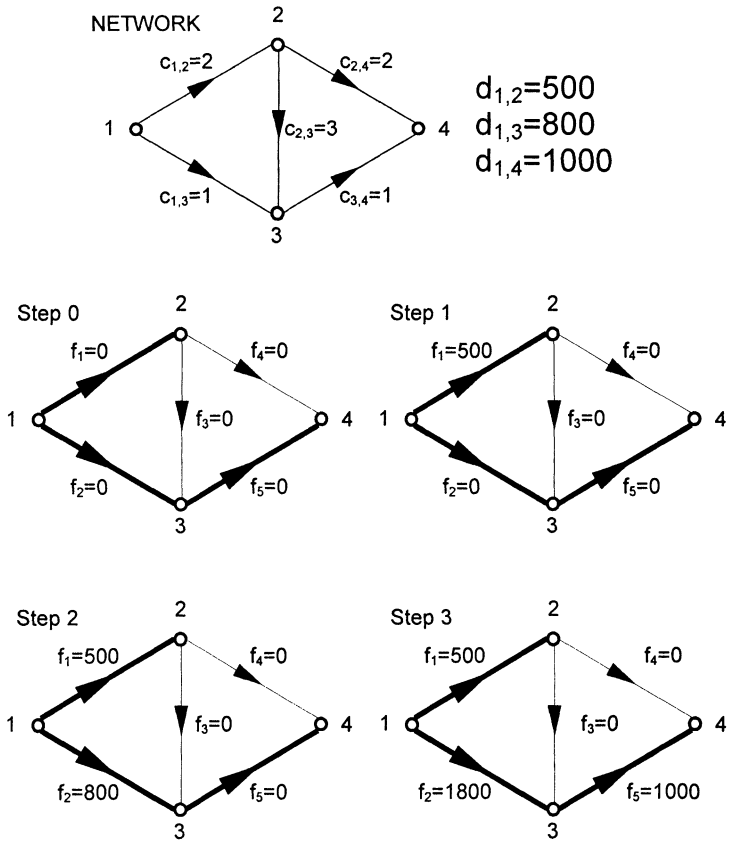


Fig. 7.3.8 Example of sequential forward algorithm for DUN assignment.

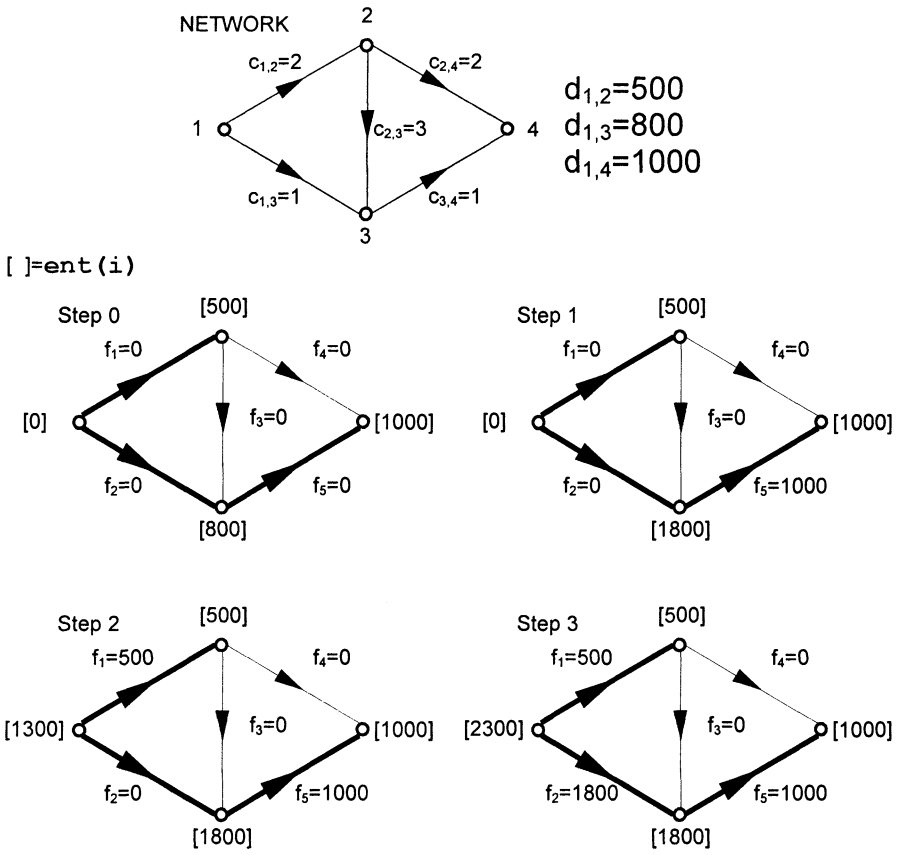


Fig. 7.3.9 Example of simultaneous forward algorithm for DUN assignment

7.4. Algorithms for rigid demand User Equilibrium assignment

In the case of congested networks, the User Equilibrium (UE) approach is usually adopted, which leads to the assignment models described in Chapter 5. This section analyzes the algorithms for solving single-class single-mode rigid demand equilibrium assignment models; some extensions will be discussed in section 7.6. In general, the algorithms for calculating equilibrium flows are based on recursive equations which, starting from a feasible solution, $\boldsymbol{f}^0 \in S_f$, generate a succession of feasible link flows vectors:

$$\boldsymbol{f}^k = \varphi(\boldsymbol{f}^{k-1}) \in S_f$$

Although the solution to the problem is not guaranteed in a finite number of steps, if the equilibrium flows vector is generated at any step k, all the remaining elements of the sequence are equal to the equilibrium vector:

$$\boldsymbol{f}^k = \boldsymbol{f}^* \Rightarrow \boldsymbol{f}^j = \boldsymbol{f}^* \quad j > k$$

Furthermore, if link flow vectors in two successive steps are equal, they are the equilibrium vector:

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} \Rightarrow \boldsymbol{f}^k = \boldsymbol{f}^*$$

Under some assumptions on cost functions and on the path choice model, it can be demonstrated that the succession defined by the recursive equations converges to the equilibrium flows vector, $\boldsymbol{f}^*$, provided that it is unique:

$$lim_{k\to\infty}\, \boldsymbol{f}^k = \boldsymbol{f}^*$$

Below, the particular case of cost functions with symmetric Jacobian is distinguished from the general case of asymmetric Jacobian, which is more difficult to solve. Recall that the algorithms described, solving the models illustrated in Section 5.4, are only those most commonly used and simplest to implement. They are essentially based on the calculation of cost functions and on the calculation of UN link flows with the algorithms described in the previous section. There are several possible variants of the basic algorithms described, or more complex algorithms based on different approaches, especially for deterministic equilibrium models (see the bibliographical note). Other algorithms for stochastic equilibrium (with separable cost functions) can be developed by solving the optimization models described in the appendix to Chapter 5; They are, however, still being researched and will not be described.

7.4.1. Rigid demand Stochastic User Equilibrium

The calculation of rigid demand stochastic user equilibrium (SUE) link flows is based on algorithms solving the fixed-point model (5.4.2) described in section 5.4.1, and repeated for ease of reference:

$$\boldsymbol{f}^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{f}^*)) \in S_f \tag{7.4.1}$$

where

$\boldsymbol{f}_{SUN}(\boldsymbol{c})$ are the Stochastic Uncongested Network (SUN) assignment link flows corresponding to link costs $\boldsymbol{c}$, calculated with the algorithms described in Section 7.3.1.

The fixed-point problem (7.4.1) can be solved with an algorithm that generates a succession of feasible link flows vectors, $\boldsymbol{f}^k$, starting from a feasible solution, $\boldsymbol{f}^0 \in S_f$. Each vector of the succession, $\boldsymbol{f}^k$, is the solution of a SUN assignment with costs corresponding to the present solution, $\boldsymbol{f}^{k-1}$. The solution of the SUN assignment is combined with the current solution, $\boldsymbol{f}^{k-1}$, to generate the next solution, $\boldsymbol{f}^k$, according to the *Method of Successive Averages* (MSA). This algorithm can be described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, e $k = 0$:

$$k = k + 1 \tag{7.4.2}$$

$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.4.3}$$

$$\boldsymbol{f}_{SUN}^{\,k} = \boldsymbol{f}_{SUN}(\boldsymbol{c}^k) \tag{7.4.4}$$

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + 1/k\,(\boldsymbol{f}_{SUN}^{\,k} - \boldsymbol{f}^{k-1}) \tag{7.4.5}$$

The solution at iteration k, $\boldsymbol{f}^k$, is the average of the first k SUN assignments; this algorithm is therefore called the *Flow Averaging* (*MSA-FA*) *algorithm*. The initial solution, $\boldsymbol{f}^0 \in S_f$, is easily obtained with a SUN assignment, using zero flow costs, $\boldsymbol{f}^0 = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{f} = \boldsymbol{0}))$. The algorithm stops if the SUN flows are equal to the current solution:

$$\boldsymbol{f}_{SUN}^{\,k} - \boldsymbol{f}^{k-1} = 0$$

Practically, the algorithm is stopped when the difference between the SUN link flows and the current solution at iteration k, $\boldsymbol{f}_{SUN}^{\,k} - \boldsymbol{f}^{k-1}$, is below a pre-assigned threshold δ, by using a suitable norm, say $|\boldsymbol{f}_{SUN}^{\,k} - \boldsymbol{f}^{k-1}| / |\boldsymbol{f}^{k-1}| < \delta$, or on a link basis, say $|f_{SUN,l}^{\,k} - f_l^{k-1}| / |f_l^{k-1}| < \delta$.

The convergence speed of the MSA algorithm close to the solution may be rather slow because the step length gets increasingly smaller. Therefore it might be convenient after a certain number of iterations to restart the algorithm with the current solution as the initial solution. This approach leads to a *two-phase algorithm* that can be generalized with a series of phases, each characterized by an increasing maximum number of iterations, for example 5 in the first phase, 10 in the second, 15 in the third, and so on.

If the cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are continuous and strictly increasing monotone, and the SUN assignment function, $\boldsymbol{f} = \boldsymbol{f}_{SUN}(\boldsymbol{c})$, is continuous and non-increasing monotone, the fixed-point problem (7.4.1) has a unique solution, as shown in section 5.4.1. Under these assumptions, by using the outcomes of Blum's theorem (see Appendix A), it can be demonstrated that the succession of feasible link flows vectors, $\boldsymbol{f}^k$, generated by the MSA-FA algorithm, converges to the equilibrium link flow vector, if the Jacobian of the cost functions is symmetric. An example of application of the MSA-FA algorithm is given in Fig. 7.4.1.

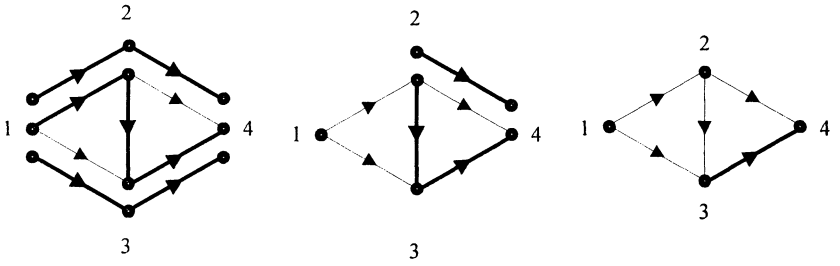
The monotonicity of the SUN assignment function is ensured if the distribution of random residuals of path choice model does not depend on the link congested cost attributes. With a Logit path choice model, this condition is ensured if the parameter θ and the definition of efficient paths are independent of the link costs $\boldsymbol{c}$ (they might depend, however, on zero flow costs, or on other cost attributes not dependent on congestion). Analogously, with a Probit path choice model this condition is ensured if the variance-covariance matrix Σ is independent of the link costs $\boldsymbol{c}$ (but, it might depend on zero flow costs or on other attributes not varying with congestion).

In the case of Probit path choice model, as was seen in Section 7.3.2.b, only an unbiased estimate of SUN flows can be obtained through a Monte Carlo algorithm. In this case almost sure convergence of the MSA-FA algorithm is assured. Furthermore, the convergence threshold δ that can be guaranteed depends on the number of iterations within the SUN assignment algorithm. To improve the overall efficiency of the SUE algorithm, initially a small number of iterations within the SUN algorithm (inner iterations) can be adopted (1 – 3) until a solution close to the equilibrium solution is obtained. In this first phase, the previous stop criterion cannot be adopted because of the small number of inner iterations, thus two successive solutions are usually compared, $\boldsymbol{f}^k \cong \boldsymbol{f}^{k-1}$, to stop the first phase. In the second phase a larger number (30 – 60) of inner iterations is adopted depending on the convergence threshold. In this second phase, the correct stop criterion can be used. Instead of this two-phase approach, a maximum number of inner iterations of the SUN algorithm increasing with the outer iteration index of the MSA algorithm can be adopted. For example, 2 iterations within the SUN algorithm for the first 10 iterations of the MSA algorithm, then 4 for the next 10, and so on.

The MSA algorithm could also be applied for SUE with non-separable cost functions; in this case, however, convergence cannot be demonstrated. A different stochastic equilibrium algorithm with non-separable cost functions (asymmetric Jacobian) can be obtained by applying the method of successive averages to costs rather than to link flows. In this case the *Cost Averaging* (*MSA-CA*) *algorithm* is obtained, specified by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $\boldsymbol{c}^0 = \boldsymbol{c}(\boldsymbol{f}^0)$, e $k = 0$ (note that the link flow vector $\boldsymbol{f}^k$ at each iteration k is feasible):

$$k = k + 1 \tag{7.4.6}$$
$$\boldsymbol{f}^k = \boldsymbol{f}_{SUN}(\boldsymbol{c}^{k-1}) \tag{7.4.7}$$
$$\boldsymbol{y}^k = \boldsymbol{c}(\boldsymbol{f}^k) \tag{7.4.8}$$
$$\boldsymbol{c}^k = \boldsymbol{c}^{k-1} + 1/k\,(\boldsymbol{y}^k - \boldsymbol{c}^{k-1}) \tag{7.4.9}$$

$d_{1\text{-}4} = 1000$ $d_{2\text{-}4} = 1500$ $d_{3\text{-}4} = 800$



PATHS

Paths	Nodes
1	1-2-4
2	1-2-3-4
3	1-3-4
4	2-3-4
5	2-4
6	3-4

PARAMETERS OF COST FUNCTIONS

Arco	c_0	a	Cap	γ
1-2	10	2	1000	4
1-3	22	2	1000	4
2-3	13	2	2500	4
2-4	20	2	1000	4
3-4	11	2	3300	4

COST FUNCTION TYPE

$$c_1 = c_0\left(1 + a\frac{f}{Cap}\right)^{\gamma}$$

LOGIT PARAMETER

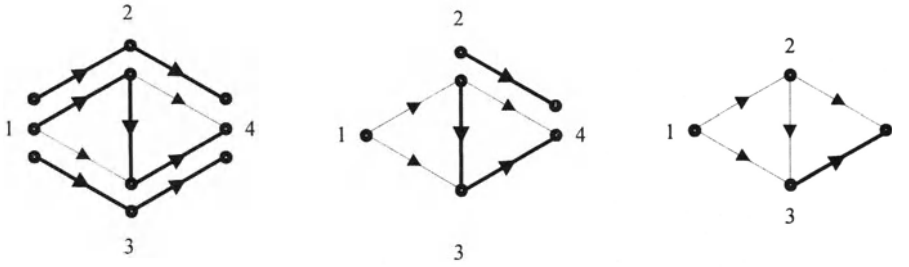
$\theta = 30$

ITERATIONS

Iteration		Link	$c(f=0)$	f^{β}_{SNL}	f^{β}
0		1-2	10	675	675
		1-3	22	325	325
		2-3	13	1060	1060
		2-4	20	1115	1115
		3-4	11	2185	2185
Iteration (k)	$\mu=1/k$	Link	$c(f^{k-1})$	f^{k}_{SNL}	f^{k}
1	1.000	1-2	14	480	480
		1-3	22	520	520
		2-3	14	1403	1403
		2-4	92	577	577
		3-4	11	2723	2723
2	0.500	1-2	11	668	574
		1-3	25	332	426
		2-3	20	1009	1206
		2-4	20	1159	868
		3-4	11	2141	2432
3	0.333	1-2	12	617	589
		1-3	23	383	411
		2-3	15	1098	1170
		2-4	32	1019	919
		3-4	11	2281	2381
4	0.250	1-2	12	596	591
		1-3	23	404	409
		2-3	15	1133	1161
		2-4	37	963	930
		3-4	11	2337	2370

Fig. 7.4.1a Example of the MSA-FA algorithm for SUE assignment (link variables).

$d_{1\text{-}4} = 1000$ $d_{2\text{-}4} = 1500$ $d_{3\text{-}4} = 800$



PATHS

Paths	Nodes
1	1-2-4
2	1-2-3-4
3	1-3-4
4	2-3-4
5	2-4
6	3-4

PARAMETERS OF COST FUNCTIONS

Arco	c_0	a	Cap	γ
1-2	10	2	1000	4
1-3	22	2	1000	4
2-3	13	2	2500	4
2-4	20	2	1000	4
3-4	11	2	3300	4

COST FUNCTION TYPE

$$c_1 = c_0\left(1 + a\frac{f}{Cap}\right)^{\gamma}$$

LOGIT PARAMETER

$\theta = 30$

FIRST ITERATIONS

Iteration		Path	$C(F=0)$	p^0	F^0_{SNL}	F^0
0		1	30	0.360	360	360
		2	34	0.315	315	315
		3	33	0.325	325	325
		4	24	0.467	700	700
		5	20	0.533	800	800
		6	11	1.000	800	800
Iteration(k)	$\mu=1/k$	Path	$C(F^{k-1})$	p^k	F^k_{SNL}	F^k
1	1.000	1	106	0.046	46	46
		2	39	0.435	435	435
		3	34	0.520	520	520
		4	25	0.905	1357	1357
		5	92	0.095	143	143
		6	11	1.000	800	800
2	0.500	1	31	0.394	394	220
		2	42	0.274	274	354
		3	36	0.332	332	426
		4	31	0.410	615	986
		5	20	0.590	885	514
		6	11	1.000	800	800
3	0.333	1	44	0.281	281	240
		2	38	0.336	336	348
		3	35	0.383	383	411
		4	26	0.545	817	930
		5	32	0.455	683	570
		6	11	1.000	800	800
4	0.250	1	50	0.352	352	254
		2	38	0.148	148	366
		3	34	0.500	500	380
		4	26	0.296	444	890
		5	37	0.704	1056	610
		6	11	1.000	800	800

Fig. 7.4.1b Example of the MSA-FA algorithm for SUE assignment (path variables).

The algorithm terminates if the SUN flows calculated with costs y^k are equal to the flows vector f^k:

$$f_{SUN}(c(f^k)) - f^k = 0$$

In practice, the algorithm terminates when the difference $f_{SUN}(c(f^k)) - f^k$ is below a pre-assigned threshold δ, by using a suitable norm or on a link basis, as above. Note that implementing the termination test is computationally demanding since a further SUN assignment has to be performed at each iteration.

The convergence of the MSA-CA algorithm is, in general, slower than that of the MSA-FA algorithm[9]. From a practical point of view, it may be convenient to perform some iterations using the MSA-FA algorithm in order to approach the equilibrium solution and then apply the MSA-CA algorithm using the current solution as the initial solution (*two-phase algorithm*). The considerations made for the MSA-FA algorithm with Probit path choice model apply also in this case.

By using the Blum theorem (see Appendix A), it can be demonstrated that the convergence of the MSA-CA algorithm is ensured if the conditions for existence and uniqueness of the solutions hold and the Jacobian of the SUN function is symmetric. Existence and uniqueness conditions require respectively continuous and strictly increasing monotone cost functions and continuous and non-decreasing monotone SUN function. The last condition is met if the distribution of random residuals in path choice model is independent of congestion dependent link cost attributes. In this case, moreover, the Jacobian of the SUN function is symmetric (as noted in section 5.3.1).

The stochastic equilibrium with non-separable cost functions can also be solved through the *inverse cost function algorithm* described in the bibliographic note. It could also be solved by applying the diagonalization algorithm, as described for deterministic equilibrium in next subsection.

7.4.2. Rigid demand Deterministic User Equilibrium

The calculation of rigid demand Deterministic User Equilibrium (DUE) link flows with symmetric Jacobian cost functions is based on algorithms solving the optimization models described in Section 5.4.2. For simplicity of notation, non-additive path costs will not be considered and the model 5.4.6b, repeated here for ease of reference, will be used:

$$f^* = argmin_{f \in S_f} z(f) = \int_0^f c(x)^T dx \qquad (7.4.10)$$

The optimization problem (7.4.10), with non-linear objective function and linear constraints, can be solved with an adaptation of the *Frank-Wolfe algorithm* (see Appendix A). This algorithm generates a succession of feasible link flows vectors, f^k, starting from a feasible solution to the problem, $f^0 \in S_f$, through the solution of a succession of linear problems approximating problem (7.4.10). The solution of the

linear problem, with respect to the current solution, $\boldsymbol{f}^{k-1}$, identifies a direction along which the objective function is minimized to determine the new solution, $\boldsymbol{f}^{k}$.

In particular, at a point $\boldsymbol{f} \in S_f$, the objective function $z(\boldsymbol{f})$ can be approximated with a linear function, $\bar{z}(\boldsymbol{f})$, using Taylor's formula up to the first term:

$$z(\boldsymbol{f}) \cong z(\bar{\boldsymbol{f}}) + \nabla z(\bar{\boldsymbol{f}})^T \boldsymbol{f} = \bar{z}(\boldsymbol{f})$$

Therefore, the optimization problem (7.4.10) can be approximated by a linear programming problem, i.e. a problem with linear objective function, $\bar{z}(\boldsymbol{f})$, and linear constraints, $\boldsymbol{f} \in S_f$:

$$argmin_{f \in Sf}\, z(\boldsymbol{f}) \cong argmin_{f \in Sf}\, \bar{z}(\boldsymbol{f}) = argmin_{f \in Sf}\, z(\bar{\boldsymbol{f}}) + \nabla z(\bar{\boldsymbol{f}})^T (\boldsymbol{f} - \bar{\boldsymbol{f}})$$

or

$$argmin_{f \in Sf}\, z(\boldsymbol{f}) \cong argmin_{f \in Sf}\, \nabla z(\bar{\boldsymbol{f}})^T \boldsymbol{f} \qquad (7.4.11)$$

Note that the gradient, $\nabla z(\boldsymbol{f})$, of the objective function, $\bar{z}(\boldsymbol{f})$, of problem (7.4.10) at a point $\bar{\boldsymbol{f}}$ is equal to the link cost vector calculated at the same point, $\nabla z(\bar{\boldsymbol{f}}) = \boldsymbol{c}(\bar{\boldsymbol{f}})$, thus expression (7.4.11) becomes:

$$argmin_{f \in Sf}\, z(\boldsymbol{f}) \cong argmin_{f \in Sf}\, \boldsymbol{c}(\bar{\boldsymbol{f}})^T \boldsymbol{f} \qquad (7.4.12)$$

The linear optimization problem expressed by (7.4.12) corresponds to the optimization model (5.3.7) described in Section 5.3.2 for Deterministic Uncongested Network assignment and can therefore be solved with a DUN algorithm as described in section 7.3.2, formally expressed by:

$\boldsymbol{f}_{DUN}(\boldsymbol{c})$ DUN link flows corresponding to a link costs vector $\boldsymbol{c}$, calculated with one of the algorithms described in Section 7.3.2.

The Frank-Wolfe algorithm for the calculation of DUE link flows with rigid demand and cost functions with symmetric Jacobian can therefore be described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $k = 0$:

$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \qquad (7.4.13)$$

$$\boldsymbol{f}_{DUN}{}^k = \boldsymbol{f}_{DUN}(\boldsymbol{c}^k) \qquad (7.4.14)$$

$$\mu^k = argmin_{\mu \in [0,1]}\, \psi(\mu) = z(\boldsymbol{f}^{k-1} + \mu (\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1})) \qquad (7.4.15)$$

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + \mu^k (\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}) \qquad (7.4.16)$$

The MSA-FA algorithm presented for stochastic equilibrium, equations (7.4.2-5), is quite similar to Frank-Wolfe. The main difference is the step, μ^k, which depends only on the iteration index, $1/k$, rather being optimized (equation 7.4.15). However, the MSA-FA algorithm may show a slower convergence.

Equation (7.4.15) defines a mono-dimensional non-linear optimization problem in the scalar variable μ that can be solved through several algorithms such as the

bisection algorithm (see Appendix A). The bisection algorithm requires the derivative of the function $\psi(\mu) = z(\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}))$, which can easily be obtained from link costs:

$$d\psi(\mu)/d\mu = \nabla z(\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}))^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}) = \\ = \boldsymbol{c}(\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}))^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1})$$

Note that in order to apply the algorithm it is not necessary to compute the value of the function $\psi(\mu)$.

From expression (7.4.16) it can be deduced that the solution at iteration k, $\boldsymbol{f}^k$, is a convex combination of the first k DUN assignments, $\boldsymbol{f}_{DUN}{}^{k}$, thus it is a feasible solution, $\boldsymbol{f}^k \in S_f$, given that the initial solution is feasible. The initial solution, $\boldsymbol{f}^0 \in S_f$, can easily be obtained, for example, with a DUN algorithm using zero flow costs, $\boldsymbol{f}^0 = \boldsymbol{f}_{DUN}(\boldsymbol{c}(\boldsymbol{f} = \boldsymbol{0}))$.

The algorithm stops when the product of the objective function gradient and the descent direction is greater than or equal to zero (see Appendix A):

$$\nabla z(\boldsymbol{f}^{k-1})^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}) = \boldsymbol{c}(\boldsymbol{f}^{k-1})^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}) \geq 0$$

It can easily be deduced that if the algorithm stops, the current solution, $\boldsymbol{f}^k$, is the DUE flow vector. In practice , the algorithm terminates when the absolute value of the product $\boldsymbol{c}(\boldsymbol{f}^{k-1})^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1})$, is below a stop threshold, δ, relative to the total cost, to avoid the effects of measurement units:

$$|\,(\boldsymbol{c}^k)^T\,(\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1})\,| \,/\, ((\boldsymbol{c}^k)^T \boldsymbol{f}^{k-1}) < \delta$$

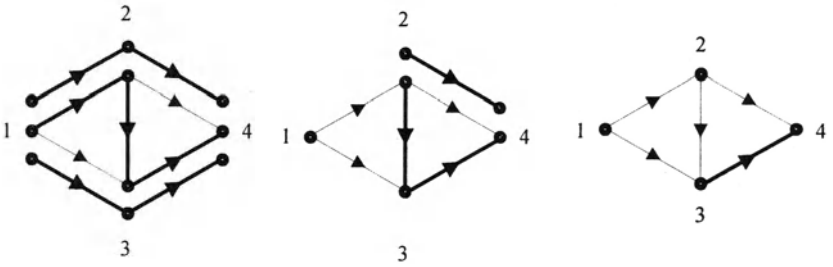
Convergence of this algorithm in proximity to the solution may be rather slow because it tends to zigzag; thus some modifications in the descent direction, $\boldsymbol{f}_{DUN}{}^{k} - \boldsymbol{f}^{k-1}$, have been proposed (some of which are referred to in Appendix A). An example of application of the Frank-Wolfe algorithm is given in Fig. 7.4.2.

If the cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$ are continuous with continuous first partial derivatives and have symmetric positive definite Jacobian, the function $z(\boldsymbol{f})$ has only one minimum point, $\boldsymbol{f}^*$, as stated in section 5.4.2. In this case also the function $\psi(\mu)$ has only one minimum point. Under these assumptions, by using results from optimization theory, it can be demonstrated that the succession of (feasible) link flows vectors, $\boldsymbol{f}^k$, generated by the Frank-Wolfe algorithm, converges to the vector of DUE link flows.

The calculation of rigid demand DUE link flows for *non-separable cost functions* (including the case of asymmetric Jacobian) is based on algorithms for solving the variational inequality model described in Section 5.4.2. For simplicity of notation, non-additive path costs are not considered and the model (5.4.4), repeated here for convenience, is used:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T\,(\boldsymbol{f} - \boldsymbol{f}^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f \tag{7.4.17}$$

$d_{1\text{-}4} = 1000$ $d_{2\text{-}4} = 1500$ $d_{3\text{-}4} = 800$



PATHS

Paths	Nodes
1	1-2-4
2	1-2-3-4
3	1-3-4
4	2-3-4
5	2-4
6	3-4

PARAMETERS OF COST FUNCTIONS

Arco	c_0	a	Cap	γ
1-2	10	2	1000	4
1-3	22	2	1000	4
2-3	13	2	2500	4
2-4	20	2	1000	4
3-4	11	2	3300	4

COST FUNCTION TYPE

$$c_1 = c_0\left(1 + a\frac{f}{Cap}\right)^{\gamma}$$

ITERATIONS

Iteration		Link	$c(f=0)$	f^0_{AoN}	f^0
0		1-2		10	1000
		1-3		22	0
		2-3		13	0
		2-4		20	2500
		3-4		11	800
Iteration (k)	**μ**	**Link**	**$c(f^{k-1})$**	**f^k_{AoN}**	**f^k**
1	0.69592	1-2	810	0	304
		1-3	22	1000	696
		2-3	13	1500	1044
		2-4	25920	0	760
		3-4	53	3300	2540
2	0.14868	1-2	67	1000	408
		1-3	720	0	592
		2-3	147	2500	1260
		2-4	807	0	647
		3-4	457	3300	2653
22	0.00006	1-2	153	1000	488
		1-3	369	0	512
		2-3	216	2500	1275
		2-4	694	0	714
		3-4	478	3300	2586
23	0.00002	1-2	153	1000	488
		1-3	369	0	512
		2-3	216	0	1275
		2-4	694	2500	714
		3-4	478	800	2586
		2-4	23	1500	679
		3-4	11	1800	2620

Fig. 7.4.2 Example of the Frank-Wolfe algorithm for DUE assignment.

Variational inequality (7.4.17) can be solved with the *diagonalization algorithm*. This algorithm generates a succession of feasible link flows vectors, $\boldsymbol{f}^k$, starting from a feasible solution to the problem, $\boldsymbol{f}^0 \in S_f$, solving a succession of separable cost functions problems that approximate the problem (7.4.17). In particular, at a solution $\boldsymbol{f} \in S_f$ the cost function of link l, $c_l(\boldsymbol{f})$, can be approximated by a separable cost function, $\bar{c}_l(f_l)$, obtained by diagonalizing the Jacobian:

$$c_l(f_1, \ldots, f_{l-1}, f_l, f_{l+1}, \ldots) \cong \bar{c}_l(\bar{f}_1, \ldots, \bar{f}_{l-1}, f_l, \bar{f}_{l+1}, \ldots) = \bar{c}_l(f_l) \qquad \forall l$$

Therefore, variational inequality (7.4.17) can be approximated by a variational inequality with separable cost functions, $\bar{c}_l(f_l)$:

$$\boldsymbol{c}(\boldsymbol{f}^*)^T(\boldsymbol{f}-\boldsymbol{f}^*) \cong \Sigma_l\, \bar{c}_l(f_l)(f_l - f_l^*) \geq 0 \quad \forall \boldsymbol{f} \in S_f \tag{7.4.18}$$

which, in turn, is equivalent to an optimization problem analogous to problem (7.4.10) described for rigid demand DUE assignment with symmetric Jacobian cost functions. Thus problem (7.4.18) can be solved as described previously. Let

$\boldsymbol{f}_{DUE}[\boldsymbol{c}(.)]$ be the (symmetric) DUE link flows corresponding to the generic link cost functions $\boldsymbol{c}(.)$ with symmetric Jacobian; $\boldsymbol{f}_{DUE}$ can be calculated, for example, with the Frank-Wolfe algorithm.

The diagonalization algorithm can be described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $k = 0$:

$$\bar{c}_l^k(f_l) = c_l(f_1^{k-1}, \ldots, f_{l-1}^{k-1}, f_l, f_{l+1}^{k-1}, \ldots) \quad \forall l \tag{7.4.19}$$

$$\boldsymbol{f}^k = \boldsymbol{f}_{DUE}[\bar{c}_l^k(f_l)] \tag{7.4.20}$$

This algorithm is therefore equivalent to performing a succession of DUE assignments with separable cost functions. These are obtained by defining a new cost function for each link where only the corresponding link flow may vary while the flows on the other links are equal to the previous equilibrium solution. The diagonalization algorithm can also be applied by averaging over the successive DUE vectors for separable cost functions, as described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $k = 0$:

$$\begin{aligned}
&k = k + 1 \\
&\bar{c}_l^k(f_l) = c_l(f_1^{k-1}, \ldots, f_{l-1}^{k-1}, f_l, f_{l+1}^{k-1}, \ldots) \qquad \forall l \\
&\boldsymbol{f}_{DUE}^k = \boldsymbol{f}_{DUE}[\bar{c}_l^k(f_l)] \\
&\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + (1/k)(\boldsymbol{f}_{DUE}^k - \boldsymbol{f}^{k-1})
\end{aligned}$$

It can easily be deduced that, in both cases, if the diagonalization converges to a solution, this is the DUE assignment with non-separable cost functions. Consistently

with the results described in section 5.4.2, if cost functions are continuous and differentiable with positive definite Jacobian, variational inequality (7.4.17) has one and only one solution. Under this assumption, the succession of link flows vectors, $\boldsymbol{f}^k$, generated by the diagonalization algorithm, converges to the equilibrium link flows vector under some conditions of the maximum value of an appropriate norm of the Jacobian matrix. In practice, to speed up the application of the algorithm, the convergence threshold of the Frank-Wolfe algorithm is decreased at each iteration of the diagonalization algorithm; alternatively symmetric deterministic equilibrium $\boldsymbol{f}_{DUE}[.]$ is heuristically substituted with a deterministic uncongested network assignment $\boldsymbol{f}_{DUN}(.)$.

7.4.3. Algorithms for System Optimal Assignment

System optimal assignment, discussed in section 5.4.4, is formulated directly through the optimization model (5.4.8) with linear constraints:

$$\boldsymbol{f}_{SO} = argmin_{\boldsymbol{f}\in S_{\boldsymbol{f}}}\ z(\boldsymbol{f}) = \boldsymbol{c}(\boldsymbol{f})^T \boldsymbol{f}$$

The Frank-Wolfe algorithm, described for symmetric deterministic equilibrium (subsection 7.4.2), can be adopted also to solve the SO problem. The algorithm is described by the following recursive equations system, given $\boldsymbol{f}^0 \in S_{\boldsymbol{f}}$:

$$\boldsymbol{g}^k = \nabla z(\boldsymbol{f}^{k-1}) = \boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f}^{k-1})]\boldsymbol{f}^{k-1} + \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.4.21}$$

$$\boldsymbol{f}_{DUN}{}^k = \boldsymbol{f}_{DUN}(\boldsymbol{g}^k) \tag{7.4.22}$$

$$\mu^k = argmin_{\mu\in[0,1]}\ \psi(\mu) = z(\boldsymbol{f}^{k-1} + \mu(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1})) \tag{7.4.23}$$

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + \mu^k(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}) \tag{7.4.24}$$

Note that unlike with deterministic equilibrium, the calculation of the gradient of the function $z(\boldsymbol{f})$, in equation (7.4.21), requires the calculation of the Jacobian of the cost functions, a task easy only for separable cost functions. Equation (7.4.23) defines the step length, μ^k, as a solution to the one-dimensional, non-linear optimization problem in the scalar variable μ. This problem can be solved with several algorithms such as the golden section algorithm (see Appendix A), which avoid the use of the derivative of the function $\psi(\mu)$ (depending on the gradient of the function $z(\boldsymbol{f})$ and therefore on the Jacobian of cost functions).

The algorithm stops if the product between the gradient of the objective function and the descent direction is greater than or equal to zero (see Appendix A):

$$\nabla z(\boldsymbol{f}^{k-1})^T(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}) = (\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1})^T(\boldsymbol{Jac}(\boldsymbol{c}(\boldsymbol{f}^{k-1}))\boldsymbol{f}^{k-1} + \boldsymbol{c}(\boldsymbol{f}^{k-1})) \geq 0$$

In order to avoid calculating the gradient of the function $z(\boldsymbol{f})$, the algorithm can terminate when the difference between the values of the function $z(\boldsymbol{f})$, in two successive iterations, is below stop threshold, δ:

$$| z(\boldsymbol{f}^k) - z(\boldsymbol{f}^{k-1}) | / z(\boldsymbol{f}^{k-1}) < \delta$$

The function $z(\boldsymbol{f})$ is strictly convex, and has a unique minimum point, if the Jacobian, $\boldsymbol{Jac}[\boldsymbol{c}(\boldsymbol{f})]$, of cost functions, $\boldsymbol{c}(\boldsymbol{f})$, is continuous and positive definite (cost functions are strictly increasing) and each link cost function, $c_l = c_l(\boldsymbol{f})$, has a Hessian matrix, $\boldsymbol{Hess}[c_l(\boldsymbol{f})]$ that is continuous and positive semi-definite (each cost function is convex). The function $\psi(\mu)$ is strictly convex if the function $z(\boldsymbol{f})$ is strictly convex. Under these assumptions, it can be demonstrated that the succession of (feasible) link flows vectors, $\boldsymbol{f}^k$, generated by the Frank-Wolfe algorithm, converges to the SO link flow vector.

7.5. Algorithms for assignment with pre-trip/en-route path choice

In this section, the algorithms described in previous sections are extended to pre-trip/en-route. Explicit reference will be made to a network representing the service provided by a public transport system, since although possible in theory, there are no examples in the literature of applications of this approach to networks representing other transportation systems, such as car or pedestrians. For transit systems, as described in sections 4.3.4.2 and 5.5, choice alternatives are travel strategies, which may be represented by hyperpaths of the service line-based network (under quite mild assumptions) rather than by paths.

Section 7.5.1 describes an extension of the shortest path algorithms to determine the shortest hyperpath without explicit enumeration of all elementary hyperpaths. This approach is particularly relevant since the number of possible hyperpaths is much larger than the number of paths and therefore explicit enumeration is particularly burdensome, if possible at all. Sections 7.5.2 and 7.5.3 describe the algorithms for Uncongested Network assignment and for rigid demand User Equilibrium assignment respectively.

7.5.1. Shortest hyperpath algorithms

The algorithms for the computation of shortest paths described in Section 7.2 can be extended to identify shortest hyperpaths, with reference to the pre-trip/en-route path choice behavior formally described in sections 4.3.4.2 and 5.5. In the following, for the sake of simplicity, it is assumed that all origins and destinations are connected, i.e. there is at least one hyperpath from each origin to each node, and from each node to each destination. For a transit system, it is assumed that the only costs due to en-route choices are waiting times at stops leading to non-additive hyperpath costs, as described in sections 5.5. Analogously to the case of shortest paths, let

$c_l = t_{mn} \geq 0$ be the cost of link $l = (m, n)$, corresponding to travel time components such as boarding, on board, alighting, and access/egress times. These attributes are associated to the corresponding types of links. (The different time components can be multiplied by appropriate homogenization coefficients not explicitly indicated here for the sake of simplicity.) As stated in Section 5.5 the waiting time associated to waiting links is not a network characteristic since it depends on the hyperpath under consideration and can be defined by the frequency of the lines as described below;

$Z_{o,d} \geq 0$ be the cost of the shortest hyperpath between the nodes o and d.

Diversion nodes, where en-route choices are made, correspond to the stops at which users choose which line to board (see sections 4.3.4.2 and 5.5). Consistent with the transit network model described in section 2.3.1, from each diversion node m there are boarding links $l = (m,n)$ connecting the different lines available at the stop, while a waiting link connects the a stop node to the diversion node m (Fig. 7.5.1). Let:

DN be the set of diversion nodes;

$pr(m)$ be the stop node preceding the diversion node m, connected by the waiting link $(pr(m), m)$;

$\varphi_{m,n} > 0$ be the frequency of the line accessed through the boarding link $l = (m, n)$, this value is associated to each boarding link in addition to the boarding time $t^s_{m,n}$ assumed constant in the following, $t^s_{m,n} = t^s$, for simplicity's sake.

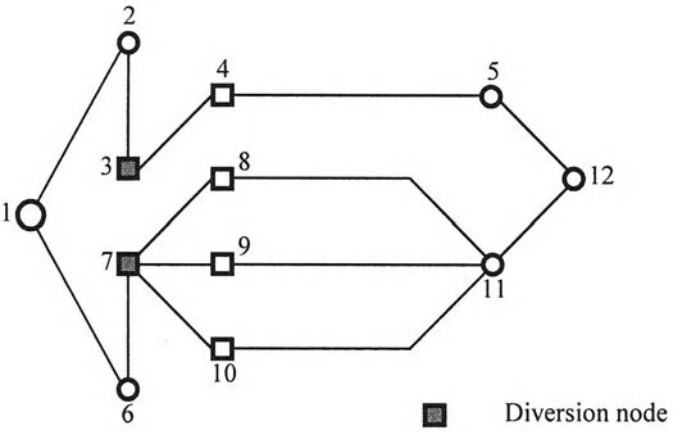


Fig. 7.5.1 Diversion nodes and adjacent elements.

The topology of a hyperpath j is identified by a succession of nodes, such that from a non-diversion node, $n \notin DN$ there may exist at most one link, while from a

diversion node, $m \in DN$, several (boarding) links, $l = (m,n)$ may exit. (Examples of hyperpaths are given in the figures in Section 4.3.4.2). When the topology of the hyperpath is known, a waiting time can be defined for each waiting link as a function of the frequencies of the lines belonging to the hyperpath considered. With reference to the hyperpath j, let:

$X^j_{m,d}$ be the cost, or travel time, from a node m to the node d along the hyperpath j;
$AL_{m,j}$ be the set of boarding links from the diversion node m in hyperpath j;
Φ^j_m be the sum of the frequencies of the lines belonging to hyperpath j and available in the diversion node m;
$t^{w,j}_m$ be the waiting time on the (unique) link $(pr(m),m)$ entering the diversion node m, in hyperpath j.

Under the assumption of random arrivals of the users, the (average) waiting time is inversely proportional to the sum of the frequencies of the lines in the hyperpath. The parameter $\theta \in [0.5, 1.0]$ depends on the regularity of the service (see section 2,5,2):

$$\Phi^j_m = (\Sigma_{(m,n)\in ALm,j}\ \varphi_{mn}) \tag{7.5.1}$$

$$t^m_{w,j} = \theta / (\Sigma_{(m,n)\in ALm,j}\ \varphi_{mn}) = \theta / \Phi^j_m \tag{7.5.2}$$

The average travel time from a diversion node m to the destination d is the average, with respect to the frequencies, of the travel times with the lines accessible from node m in hyperpath j (as noted in section 5.5):

$$X^j_{m,d} = \Sigma_{(m,n)\in ALm,j}\ (t^s + X^j_{n,d}\ (\varphi_{mn} / \Phi^j_m) \tag{7.5.3}$$

The average travel time, $X^j_{pr(m),d}$, to reach the destination d from the stop node $pr(m)$ connected to the waiting node m can be defined as the sum of the time from the diversion node, $X^j_{m,d}$,, and the waiting time, $t^{w,j}_m$:

$$X^j_{pr(m),d} = X^j_{m,d} + t^{w,j}_m \tag{7.5.4}$$

The relation (7.5.3) allows one to express the average minimum travel time from a diversion node m to the destination d, $Z_{m,d}$, as the average, with respect to the frequencies, of the minimum times along the lines from the node m belonging to the shortest hyperpath j^*:

$$Z_{m,d} = \Sigma_{(mn)\in ALm,j^*}\ (t^s + Z_{m,d})\ (\varphi_{m,n} / \Phi^{j^*}_m)$$

The shortest travel time, $Z_{pr(m),d}$, from the stop node $pr(m)$ connected to the diversion node m can be obtained by summing the shortest travel time from the diversion node, $Z_{m,d}$, and the waiting time $t^{w,j}_m{}^*$:

$$Z_{pr(m),d} = Z_{i,d} + t^{w,j}{}_m *$$

(to be compared with the shortest travel time along the access network). All the above relations can be used to extend the Bellmann principle to the shortest hyperpath problem.

It should be noted that if the forward shortest hyperpath tree from an origin o to all the other nodes were searched, it would be necessary to differentiate, at each stop, the users by destination, to take account of the different lines available. For this reason, it is useful to adopt algorithms based on an extension of the *backward updating step*, previously defined for shortest paths, allowing the determination of the tree of the shortest hyperpaths from all the nodes towards the destination d, $T(d)$.

Now consider a hyperpath j (not necessarily the shortest one) which does not include the boarding link (m,n), see Fig. 7.5.1.When examining a line node n the backward updating step must be extended to check whether the inclusion of the boarding link $l = (m,n)$ in the hyperpath j can ensure a reduction of the average travel time from node $pr(m)$. Let

$\Phi^j{}_m$, $t^{w,j}{}_m$ be the values of the cumulative frequency and the average waiting time relative to this hyperpath as defined by (7.5.1) and (7.5.2).

The average travel times from the waiting node m and the stop node $pr(m)$ in hyperpath j are given by (7.5.3), and (7.5.4), respectively. Note that the node $pr(m)$ might be connected to the destination d through other paths using the access links.

Moreover, in what follows it will be assumed that an *algorithm with ordering* will be adopted, for reasons that will become clear below. Thus, let:

$Z_{n,d}$ be the minimum cost, or travel time, between the line node n and destination d, already known when node n is examined.

If the boarding link (m,n) is added to the hyperpath j, a further line with frequency $\varphi_{m,n}$ is available at the stop node m. Therefore, there is a further path to reach the destination d. The new hyperpath j' induces a reduction of the average travel time from the node $pr(m)$ to the destination d, if it results:

$$X^{j'}{}_{pr(m),d} \leq X^{j}{}_{pr(m),d} \tag{7.5.5}$$

To analyze the implications of (7.5.5), note that the hyperpath j' has a larger cumulative frequency at node m and a smaller waiting time, obtained by applying the relationships (7.5.1) and (7.5.2):

$$\Phi^{j'}{}_m = \Phi^{j}{}_m + \varphi_{mn} \tag{7.5.6}$$

$$t^{w,j'}{}_m = \theta / \Phi^{j'}{}_m = t^{w,j}{}_m \left[\Phi^{j}{}_m / (\Phi^{j}{}_m + \varphi_{mn})\right] \tag{7.5.7}$$

The inclusion of the second line causes also a variation of the average travel time from the node m to the destination d that from (7.5.3) becomes:

$$X^{j'}_{m,d} = X^{j}_{m,d}[\Phi^{j}_{m} / (\Phi^{j}_{m} + \varphi_{mn})] + (Z_{n,d} + t^{s}_{mn})\,[\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn})] =$$
$$= X^{j}_{m,d} + (Z_{n,d} + t^{s}_{mn} - X^{j}_{m,d})\,\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn}) \qquad (7.5.8)$$

since $[\Phi^{j}_{m} / (\Phi^{j}_{m} + \varphi_{mn})] = 1 - [\varphi_{ij} / (\Phi^{j}_{m} + \varphi_{ij})]$. Thus, the average travel time from the waiting node $pr(m)$ to the destination d, through diversion node m, after the introduction of the boarding link (m,n), becomes:

$$X^{j'}_{pr(m),d} = X^{j'}_{m,d} + t^{w,j}_{m} \qquad (7.5.9)$$

or

$$X^{j'}_{pr(m),d} = X^{j}_{m,d} + (Z_{n,d} + t^{s}_{mn} - X^{j}_{m,d})\,\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn}) + t^{w,j}_{m}\,[\Phi^{j}_{m} / (\Phi^{j}_{m} + \varphi_{mn})]$$

The above relationship combined with the condition (7.5.5) becomes:

$$X^{j}_{m,d} + (Z_{n,d} + t^{s}_{mn} - X^{j}_{m,d})\,\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn}) + t^{w,j}_{m}\,[\Phi^{j}_{m} / (\Phi^{j}_{m} + \varphi_{mn})] \le X^{j}_{m,d} + t^{w,j}_{m}$$
$$(Z_{n,d} + t^{s}_{mn} - X^{j}_{m,d})\,\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn}) \le t^{w,j}_{m}\,\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn})$$
$$Z_{n,d} + t^{s}_{mn} \le X^{j}_{m,d} + t^{w,j}_{m}$$

since $[\varphi_{mn} / (\Phi^{j}_{m} + \varphi_{mn})] > 0$. Therefore it is worth including link $l = (m,n)$ if:

$$Z_{n,d} + t^{s}_{mn} \le X^{j}_{pr(m),d} \qquad (7.5.10)$$

On the other hand, given the hyperpath j' containing the boarding link $l = (m,n)$, it is not possible to obtain a reduction of the total travel time excluding this link $l = (m,n)$ from the hyperpath if condition (7.5.10) is verified, (vice versa if the condition is not verified).

Condition (7.5.10) shows that, to reduce the average travel time and find the shortest hyperpath, it is useful to include a new line if the travel time with the new line, including boarding time, is less than the travel time, including waiting time, without the line. If this is the case, the inclusion of the new line reduces the waiting time so that even if the average travel time from the diversion node increases, the average travel time from the stop node decreases.

The shortest hyperpath for a pair (o,d) may not include any waiting links (and therefore boarding, line and alighting links). In this case it consists only of access links, implying that the shortest path on the access network has a cost lower than any paths using a transit line.

The algorithms for calculating the tree of shortest hyperpaths towards a destination d are similar to those described in Section 7.2 for the shortest paths. The main difference is the updating step that also includes the updating operations of the tentative minimum cost value of a diversion node, using condition (7.5.10) and relations (7.5.8) and (7.5.9) to update average travel times. In this way a stop node might be connected to the destination d by other paths through the access links adjacent to it. The tree of the shortest hyperpaths towards the destination node d,

T(d), can be described by the link, necessarily unique, exiting from each node, except for diversion nodes from which there might be several boarding links, identifying the lines included in the shortest hyperpath.

Note that the node made definitive at each iteration should be the one with the minimum tentative cost among non-definitive nodes and the updating step should be performed from this node. Therefore, for the identification of the tree of the shortest paths towards the destination *d*, *T(d), algorithms with ordering* (in the sense defined previously for the shortest paths) should be adopted. In addition, consider a further boarding link (m,r) not included in the hyperpath j such that $Z_{r,d} \leq Z_{n,d}$, or $Z_{r,d} + t^s \leq Z_{n,d} + t^s$ if the condition (7.5.10) is verified for link (m,r), and therefore it is convenient to include link (m,n) to reduce the average cost, it is also verified for link (m,n), and it is even more convenient to include also link (m,r). This consideration further supports the adoption of algorithms with ordering, in which the updating of the line nodes n connected to a diversion node m through the boarding links (m,n) is carried out by increasing values of $Z_{n,d}$. Otherwise it would be necessary to verify, at each new inclusion, if some of the boarding links already included should be removed. Algorithms with ordering terminate after as many updating steps as there are nodes, since at each step the tentative value of a node is made definitive. Nodes are made definitive in order of increasing minimum costs, or travel times, to the destination. At the end of the algorithm, the waiting times, specific to the shortest hyperpaths, and the set of boarding links for each diversion node, are also determined.

7.5.2. Algorithms for Uncongested Network assignment with pre-trip/en-route path choice

Uncongested Network assignment models with pre-trip/en-route path choice are quite often adopted for analyzing public transportation systems assuming, as a first approximation, that the costs are not dependent on users flows. Furthermore, algorithms for UN assignment are included in equilibrium assignment algorithms as described in the following section. In the case of explicit hyperpaths enumeration calculation of link flows is straightforward by using the sequence of relations given in Section 5.5. In general, however, as already noted, explicit enumeration of the hyperpaths is extremely burdensome and algorithms without explicit enumeration based on the shortest hyperpaths algorithms previously described are adopted.

Stochastic Uncongested Network assignment algorithms, with Probit choice models can easily be extended to transit networks, provided that All-or-Nothing algorithms are extended as described below. They essentially require multiple sampling of perceived link costs (and possibly frequencies) as in the Monte Carlo algorithm described in section 7.3.1.b. However, very few examples are described in the literature. Generalization to the case of Logit hyperpath choice without explicit hyperpaths enumeration is still at the research stage (see bibliographical note).

On the other hand, under the assumption of deterministic choice behavior, all the users traveling between each O-D pair choose the shortest hyperpath (Section 5.5).

If there are several shortest hyperpaths for some O-D pair, hyperpaths flows, and therefore links flows, are not uniquely defined. Also in this case, however, the shortest hyperpaths tree algorithm gives a unique hyperpath between each pair. Link flows can be calculated by assigning the demand flow for each O-D pair to the links of the shortest hyperpath and summing for all O-D pairs.

The backward simultaneous algorithm discussed for DUN assignment in section 7.3.2, can be extended to the case of shortest hyperpaths. Also in this case an algorithm with ordering is required for shortest hyperpath trees to each destination. Operations performed at a diversion node must be modified. In this case the exit flow must be divided among all boarding links included in the hyperpaths tree, proportionally to their frequencies. The application of DUN algorithms to shortest hyper paths yields the links flows f_{DUN} as a function of the costs of non-waiting links, c, and of the frequencies of the lines, φ. It is also possible to calculate the hyperpath total non-additive cost, $X^{NA}{}_{DUN}$, given by the total waiting time, which can be determined with shortest hyperpath algorithms without explicit enumeration.

7.5.3. Algorithms for rigid demand User Equilibrium assignment with pre-trip/en-route path choice

The algorithms described in Section 7.4 for rigid demand stochastic or deterministic equilibrium assignment can be extended to pre-trip/en-route path choice. The main modification occurs in the calculation of the UN flows with the procedure described in the previous section. Furthermore, it is necessary to consider explicitly the non-additive hyperpath costs consisting of the waiting times. Let

$\boldsymbol{x}^{NA}$ be vector of non-additive hyperpath costs, consisting of the vectors of non-additive hyperpath costs $\boldsymbol{x}_{od}^{NA}$ for each pair od (assumed to be independent of congestion);

$X^{NA} = (\boldsymbol{x}^{NA})^T \boldsymbol{y}$ be total non-additive cost, i.e. waiting time, corresponding to the generic vector of hyperpath flows $\boldsymbol{y}$.

In the case of stochastic equilibrium, a fixed-point problem similar to problem (7.4.1) in terms of link flows is obtained:

$$\boldsymbol{f}^* = \boldsymbol{f}_{SUN}(\boldsymbol{c}(\boldsymbol{f}^*); \boldsymbol{d}) = \sum_{od} d_{od} \boldsymbol{\Lambda}_{od} \boldsymbol{p}_{j,od}(-(\boldsymbol{\Lambda}_{od}^T \boldsymbol{c}(\boldsymbol{f}^*) + \boldsymbol{x}_{od}^{NA})) \in S_f \qquad (7.5.11)$$

Model (7.5.11) can be solved with the MSA-FA and MSA-CA algorithms already described in section 7.4.1, implementing at each iteration a Stochastic Uncongested Network assignment to the hyperpaths.

In the case of symmetric deterministic equilibrium, the model (7.4.10) becomes:

$$(\boldsymbol{f}^*, X^{NA}*) = argmin\, z(\boldsymbol{f}, X) = \int_0^{\boldsymbol{f}} \boldsymbol{c}(\boldsymbol{x})^T d\boldsymbol{x} + X^{NA} \qquad (7.5.11)$$
$$\boldsymbol{f} = \boldsymbol{\Lambda} \boldsymbol{y},\ X^{NA} = (\boldsymbol{x}^{NA})^T \boldsymbol{y},\quad \boldsymbol{y} \in S_y$$

This model can be solved with the Frank-Wolfe algorithm, described in section 7.4.2 considering as variables the link flows vector, $\boldsymbol{f}$, and the non-additive hyperpath total cost (total waiting time), X^{NA}. The following variables are needed to describe the algorithm:

$\nabla z(\boldsymbol{f},X^{NA})^T = [\boldsymbol{c}(\boldsymbol{x}), 1]$	gradient of the function $z(\boldsymbol{f},X^{NA})$;
$\boldsymbol{f}_{DUN} \in S_f$	link flows resulting from DUN assignment to hyperpaths as function of the total costs on non-waiting links, $\boldsymbol{c}$, and the lines frequences, φ;
$X^{NA}{}_{DUN}$	total non-additive hyperpath cost equal to the waiting time, resulting from the non-additive hyperpaths assignment as a function of non-waiting link costs $\boldsymbol{c}$, and of lines frequencies, φ;
$(\boldsymbol{f}_{DUN}, X^{NA}{}_{DUN}) = \text{DUN}(\boldsymbol{c}, \varphi)$	a function giving $\boldsymbol{f}_{DUN}$ and $X^{NA}{}_{DUN}$ as a function of $\boldsymbol{c}$, and φ.

Given an initial solution, $(\boldsymbol{f}^0, X^{NA0})$, that can easily be found with a DUN assignment algorithm using zero flow costs, $(\boldsymbol{f}^0, X^{NA0}) = \text{DUN}(\boldsymbol{c}(\boldsymbol{f}= \boldsymbol{0}), \varphi)$, the Frank-Wolfe algorithm for the solution of the model (7.5.11) can be described by the system of following recursive equations:

$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.5.12}$$

$$(\boldsymbol{f}_{DUN}{}^k, X^{NA}{}_{DUN}{}^k) = \text{DUN}(\boldsymbol{c}^k, \varphi) \tag{7.5.13}$$

$$\mu^k = argmin_{\mu\in[0,1]}\ \psi(\mu) = \\ = z((\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1})), (X^{NA\,k-1} + \mu\,(X^{NA}{}_{DUN}{}^k - X^{NA\,k-1}))) \tag{7.5.14}$$

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + \mu^k\,(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}) \tag{7.5.15}$$

$$X^{NA\,k} = X^{NA\,k-1} + \mu^k\,(X^{NA}{}_{DUN}{}^k - X^{NA\,k-1}) \tag{7.5.16}$$

Equation (7.5.14) defines a mono-dimensional non-linear optimization problem in the scalar variable μ that can be solved with several algorithms, such as the bisection algorithm (see Appendix A). This algorithm uses the derivative of the objective function $\psi(\mu)$, which can be easily computed from link costs:

$$\begin{aligned} d\psi(\mu)/d\mu = \\ &= \nabla z[(\boldsymbol{f}^{k-1}+\mu\,(\boldsymbol{f}_{DUN}{}^k-\boldsymbol{f}^{k-1})), (X^{NA\,k-1}+\mu\,(X^{NA}{}_{DUN}{}^k-X^{NA\,k-1}))]^T \cdot \\ &\quad \cdot[(\boldsymbol{f}_{DUN}{}^k-\boldsymbol{f}^{k-1}), (X^{NA}{}_{DUN}{}^k-X^{NA\,k-1})] = \\ &= [\boldsymbol{c}(\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1})) , 1]^T\ [(\boldsymbol{f}_{DUN}{}^k-\boldsymbol{f}^{k-1}), (X^{NA}{}_{DUN}{}^k-X^{NA\,k-1})] = \\ &= \boldsymbol{c}(\boldsymbol{f}^{k-1} + \mu\,(\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}))^T\ (\boldsymbol{f}_{DUN}{}^k - \boldsymbol{f}^{k-1}) + (X^{NA}{}_{DUN}{}^k-X^{NA\,k-1}) \end{aligned}$$

Note that the algorithm does not require calculation of the function $\psi(\mu)$.

If cost functions $c = c(f)$ are continuous with continuous first partial derivatives and with positive definite symmetric Jacobian, the term $\int_0^f c(v)^T\, dv$ is a strictly convex function of f. In this case the function $z(f, X^{NA})$ has one and only one minimum point $(f^*, X^{NA}*)$ as already seen in section 5.5. In this case, the function $\psi(\mu)$ has one and only one minimum point. Under these assumptions, from the results of optimization theory, it can be demonstrated that the succession of (feasible) link flows vectors, f^k , generated by the Frank-Wolfe algorithm, converges to the deterministic equilibrium link flow vector, and so the values $X^{NA,k}$.

Deterministic equilibrium with non-separable cost functions can be analyzed with variational inequality models in terms of link flows f^* and of total non-additive cost $X^{NA}*$. This problem can be formalized as:

$$c(f^*)^T(f - f^*) + (X^{NA} - X^{NA}*) \geq 0 \quad \forall f = \Lambda y, \forall X^{NA} = (x^{NA})^T y,\ \forall y \in S_y \qquad (7.5.17)$$

and solved with the diagonalization algorithm discussed in section 7.4.2.

7.6. Extensions of User Equilibrium assignment algorithms*

This section briefly describes extensions of the rigid demand equilibrium assignment algorithms presented in section 7.4 to the case of elastic demand equilibrium assignment. The algorithms described can easily be adapted to solve multi-mode equilibrium assignment, which will not be discussed below in details.

As seen in section 5.6, elastic demand assignment models assume that O-D demand flows depend on congested transportation costs. This assumption implies that users' behavior on choice dimensions other than the path (e.g. mode, destination) is influenced by variations of path costs due to variations of congestion levels. In single-mode assignment, it is assumed that costs of only one mode depend on congestion. In this case, the dependence of demand flows on path costs can be expressed by demand functions, which depend on the EMPU function relative to the path choice model, see section 5.6:

$$s = s(V = -g)$$
$$d = d(s)$$

Calculation of link and demand flows for elastic demand (single-mode) equilibrium assignment can be performed with three different approaches described below.

External cycle algorithms solve a formulation of elastic demand equilibrium assignment models in which the circular dependence between demand flows and costs is expressed externally to the flow-cost equilibrium. As stated in section 5.6, this defines a two-level problem. Equilibrium between flows and costs is computed at the inner level for given demand. At the outer level equilibrium between costs resulting from the equilibrium assignment and demand flows resulting from demand functions is computed. Let

$\boldsymbol{f}_{UE\text{-}RIG} = \boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d})$ be the implicit correspondence, between rigid demand equilibrium link flows, $\boldsymbol{f}_{UE\text{-}RIG}$, and demand flows $\boldsymbol{d}$. This correspondence expresses the solution of one of the models described in Section 5.4. If the equilibrium link flow vector is unique for a given demand vector, the above correspondence is a one-to-one function. Its value can be calculated with one of the algorithms described in Section 7.4.

Elastic demand equilibrium assignment can be formulated with a system of non-linear equations:

$$\boldsymbol{d}^* = \boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}(\boldsymbol{f}^*))) \tag{7.6.1}$$
$$\boldsymbol{f}^* = \boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}^*) \tag{7.6.2}$$

Combining the two equations (7.6.1) and (7.6.2), a combined fixed-point problem (with an implicitly defined function) in the demand flows, $\boldsymbol{d}^*$, or in the link flows, $\boldsymbol{f}^*$, is obtained:

$$\boldsymbol{d}^* = \boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}(\boldsymbol{f}_{UE}(\boldsymbol{d}^*)))) \tag{7.6.3}$$
$$\boldsymbol{f}^* = \boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}(\boldsymbol{f}^*)))) \tag{7.6.4}$$

The fixed-point problem can also be formulated in link costs or in EMPU values.

The simplest external cycle algorithms are based on the iterative application of a rigid demand equilibrium assignment algorithm, for the calculation of link flows and costs with given demand flows and of the demand function for the calculation of demand flows with given costs and EMPU's. In particular, an external cycle algorithm of this type can be specified by the following system of recursive equations, given an initial value of the demand flows, $\boldsymbol{d}^0 \in S_d$:

$$\boldsymbol{f}^k = \boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}^{k-1}) \tag{7.6.5}$$
$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^k) \tag{7.6.6}$$
$$\boldsymbol{s}^k = \boldsymbol{s}(-\Delta^T \boldsymbol{c}^k) \tag{7.6.7}$$
$$\boldsymbol{d}^k = \boldsymbol{d}(\boldsymbol{s}^k) \tag{7.6.8}$$

The initial value of the demand flows $\boldsymbol{d}^0$ can be obtained with EMPU's corresponding zero flow link costs: $\boldsymbol{c}^0 = \boldsymbol{c}(\boldsymbol{f} = \boldsymbol{0})$, $\boldsymbol{s}^0 = \boldsymbol{s}(-\Delta^T \boldsymbol{c}^0)$, $\boldsymbol{d}^0 = \boldsymbol{d}(\boldsymbol{s}^0)$.

A more sophisticated external cycle algorithm can be specified by applying the MSA to the fixed-point problem (7.6.3) in demand flows, $\boldsymbol{d}^*$. The resulting algorithm can be described by the following system of recursive equations, given $\boldsymbol{d}^0 \in S_d$, and $k = 0$:

$$k = k - 1 \tag{7.6.9}$$
$$\boldsymbol{f}^k = \boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}^{k-1}) \tag{7.6.10}$$
$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^k) \tag{7.6.11}$$
$$\boldsymbol{s}^k = \boldsymbol{s}(-\Delta^T \boldsymbol{c}^k) \tag{7.6.12}$$

$$\boldsymbol{d}^k = \boldsymbol{d}^{k-1} + (1/k)\,(\boldsymbol{d}(\boldsymbol{s}^k) - \boldsymbol{d}^{k-1}) \tag{7.6.13}$$

Analogously, an external cycle algorithm can be specified by applying the MSA method to the fixed-point problem (7.6.4) in link flows. This produces an algorithm described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, and $k = 0$:

$$k = k - 1 \tag{7.6.14}$$
$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.6.15}$$
$$\boldsymbol{s}^k = \boldsymbol{s}(-\Delta^T \boldsymbol{c}^k) \tag{7.6.16}$$
$$\boldsymbol{d}^k = \boldsymbol{d}(\boldsymbol{s}^k) \tag{7.6.17}$$
$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + (1/k)\,(\boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}^k) - \boldsymbol{f}^{k-1}) \tag{7.6.18}$$

In both cases termination tests should compare the value at the previous iteration, ($\boldsymbol{d}^{k-1}$ or $\boldsymbol{f}^{k-1}$) with the value obtained within the iteration ($\boldsymbol{d}(\boldsymbol{s}^k)$ or $\boldsymbol{f}_{UE\text{-}RIG}(\boldsymbol{d}^*)$).

Other algorithms can be specified by applying the MSA method to the EMPU's, to link costs, or to pairs of variables. It is easily deduced that, whatever the case, if an external cycle algorithm converges at a solution, this is the equilibrium solution sought. The convergence of external algorithms has not yet been completely analyzed nor have the conditions on assignment models and on demand functions that ensure it. External algorithms are easily implemented, starting from existing rigid demand assignment implementations and particularly flexible for variations of the demand functions.

Internal cycle algorithms are based on extension of the algorithms solving rigid demand equilibrium assignment problems described in Section 7.4. In the case of elastic demand equilibrium, it is rather straightforward to extend the MSA-FA or MSA-CA algorithms described for rigid demand stochastic equilibrium (section 7.4.). At each iteration, these algorithms compute the EMPU's and therefore demand flows with costs at the previous iteration, before proceeding to the UN assignment of that demand. This approach is simple to apply with or without explicit path enumeration.

In the case of Logit SUN (without explicit paths enumeration), the Dial algorithm described in Section 7.3.1.a can easily be extended. In particular, for each origin o, after the calculation of the weights for the nodes, W_i, and for the links, w_{ij}, in the first phase of the algorithm, the inclusive variable, Y_d, is obtained for each destination d. This variable is the EMPU, s_{od}, between the pair od. Demand flow, d_{od}, can thus be computed and loaded on the network with the Dial algorithm.

In the case of Probit SUN (without explicit paths enumeration), the Monte Carlo algorithm described in Section 7.3.1.b can be extended quite easily. In particular, for each pair od the average of the shortest path costs corresponding to the sampled perceived costs is an unbiased estimate of the opposite of the EMPU, $\bar{s}_{od}$. From these estimates d_{od} flows can be estimated and, from them, link flows:

$$\bar{\boldsymbol{s}}^m = \bar{\boldsymbol{s}}^m(\boldsymbol{c})$$
$$\bar{\boldsymbol{d}} = \boldsymbol{d}(\bar{\boldsymbol{s}})$$

$$\bar{\boldsymbol{f}}^m = \bar{\boldsymbol{f}}^m(\boldsymbol{c}, \boldsymbol{d})$$

where

$\bar{\boldsymbol{s}}^m = \bar{\boldsymbol{s}}^m(\boldsymbol{c})$ is a vector of un-biased estimates of the EMPU's for all pairs od, obtained with a sample of m perceived link costs with mean $\boldsymbol{c}$;

$\bar{\boldsymbol{f}}^m = \bar{\boldsymbol{f}}^m(\boldsymbol{c}, \boldsymbol{d})$ is a un-biased estimate of SUN link flows resulting from demand flows $\boldsymbol{d}$ and a sample of m vectors of perceived link costs with mean $\boldsymbol{c}$.

Note that the direct application of this approach, given a vector $\boldsymbol{c}$, requires two repetitions of the estimation process, first for the EMPU's and then for links flows.

In the case of deterministic UN assignment (without explicit paths enumeration), the algorithms described in Section 7.3.2 can easily be extended. In particular, for each origin o, the algorithm for determining the shortest paths tree for each origin o gives the minimum cost, Z_{od}, between o and all destinations d. The opposite of these values are the EMPU's, $s_{od} = -Z_{od}$, from which demand flows d_{od} can be computed and assigned to the links of the shortest path between o and d.

Whatever procedure is adopted for UN assignment, either stochastic or deterministic, with or without explicit paths enumeration, the MSA-FA algorithm for internal cycle elastic demand equilibrium can be defined by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $\boldsymbol{d}^0 \in S_d$, and $k = 0$:

$$k = k + 1 \tag{7.6.19}$$

$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.6.20}$$

$$\boldsymbol{f}_{UN}{}^k = \boldsymbol{f}_{UN}(\boldsymbol{c}^k, \boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}^k))) \tag{7.6.21}$$

$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + 1/k\,(\boldsymbol{f}_{UN}{}^k - \boldsymbol{f}^{k-1}) \tag{7.6.22}$$

where

$\boldsymbol{f}_{UN}(\boldsymbol{c}, \boldsymbol{d})$ are the link flows resulting from a UN assignment with costs $\boldsymbol{c}$ and demand flows $\boldsymbol{d}$;

$\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}))$ are the demand flows corresponding to the EMPUs relative to link costs $\boldsymbol{c}$.

Note the difference between the external cycle algorithm (eqns. 7.6.9-7.6.13) and internal (eqns. 7.6.19-7-6-22). In the first case, at each iteration a rigid demand equilibrium assignment is performed, requiring several UN assignments; then resulting link flows are averaged. Vice versa, in the internal cycle algorithm at each iteration only one UN assignment is performed and resulting link flows are averaged. There are no systematic comparisons of the two approaches. From the purely computational point of view the relative efficiency is certainly related to the relative complexity of computing UN flows and demand flows.

The internal cycle MSA-FA algorithm can be further extended averaging EMPU values as well as the link flows, as described by following system of the recursive equations, given $\boldsymbol{f}^0 \in S_f$, $\boldsymbol{s}^0 = \boldsymbol{s}(-\Delta^T \boldsymbol{c}(\boldsymbol{f}^0))$ e $k = 0$:

$$k = k + 1 \tag{7.6.23}$$
$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{f}^{k-1}) \tag{7.6.24}$$
$$\boldsymbol{d}^k = \boldsymbol{d}(\boldsymbol{s}^{k-1}) \tag{7.6.25}$$
$$(\boldsymbol{s}_{UN}^k, \boldsymbol{f}_{UN}^k) = \boldsymbol{UN}(\boldsymbol{c}^k, \boldsymbol{d}^{kj}) \tag{7.6.26}$$
$$\boldsymbol{s}^k = \boldsymbol{s}^{k-1} + 1/k\,(\boldsymbol{s}_{UN}^k - \boldsymbol{s}^{k-1}) \tag{7.6.27}$$
$$\boldsymbol{f}^k = \boldsymbol{f}^{k-1} + 1/k\,(\boldsymbol{f}_{UN}^k - \boldsymbol{f}^{k-1}) \tag{7.6.28}$$

where

$(\boldsymbol{s}_{UN}, \boldsymbol{f}_{UN}) = \boldsymbol{UN}(\boldsymbol{c}, \boldsymbol{d})$ are the EMPU and flows resulting from UN assignment with link costs $\boldsymbol{c}$ and demand flows $\boldsymbol{d}$; they can be computed simultaneously with one of the procedures described in section 7.3.

This algorithm, called MSA-FSA, is particularly useful in the case of Probit path choice model since it avoids the double Monte Carlo application at each iteration.

In the case of equilibrium with non-separable cost functions (asymmetric Jacobian), the convergence of the MSA-FA and MSA-FSA algorithms has not been proved. It is possible to adopt an immediate extension of the MSA-CA algorithm or the diagonalization algorithm (described in Section 7.4 for rigid demand equilibrium). In particular, the MSA-CA algorithm can be described by the following system of recursive equations, given $\boldsymbol{f}^0 \in S_f$, $\boldsymbol{c}^0 = \boldsymbol{c}(\boldsymbol{f}^0)$, and $k = 0$:

$$k = k + 1 \tag{7.6.29}$$
$$\boldsymbol{f}_{UN}^k = \boldsymbol{f}_{UN}(\boldsymbol{c}^{k-1}, \boldsymbol{d}(\boldsymbol{s}(-\Delta^T \boldsymbol{c}^{k-1}))) \tag{7.6.30}$$
$$\bar{\boldsymbol{c}}^k = \boldsymbol{c}(\boldsymbol{f}^k) \tag{7.6.31}$$
$$\boldsymbol{c}^k = \boldsymbol{c}^{k-1} + 1/k\,(\bar{\boldsymbol{c}}^k - \boldsymbol{c}^{k-1}) \tag{7.6.32}$$

Note that the link flows vector $\boldsymbol{f}^k = \boldsymbol{f}_{UN}(\boldsymbol{c}^{k-1})$ at the iteration k is feasible.

In general, it is possible to average both demand flows, and link costs, with an algorithm called MSA-CDA, described by the system of following recursive equations, given $\boldsymbol{f}^0 \in S_f$, $\boldsymbol{d}^0 \in S_d$, $\boldsymbol{c}^0 = \boldsymbol{c}(\boldsymbol{f}^0)$, and $k = 0$:

$$k = k + 1 \tag{7.6.33}$$
$$\boldsymbol{f}^k = \boldsymbol{f}_{UN}(\boldsymbol{c}^{k-1}, \boldsymbol{d}^{k-1}) \tag{7.6.34}$$
$$\bar{\boldsymbol{d}}^k = \boldsymbol{d}(\boldsymbol{s}(-\boldsymbol{A}^T \boldsymbol{c}^k)) \tag{7.6.35}$$
$$\bar{\boldsymbol{c}}^k = \boldsymbol{c}(\boldsymbol{f}^k) \tag{7.6.36}$$
$$\boldsymbol{d}^k = \boldsymbol{d}^{k-1} + 1/k\,(\bar{\boldsymbol{d}}^k - \boldsymbol{d}^{k-1}) \tag{7.6.37}$$
$$\boldsymbol{c}^k = \boldsymbol{c}^{k-1} + 1/k\,(\bar{\boldsymbol{c}}^k - \boldsymbol{c}^{k-1}) \tag{7.6.38}$$

The convergence of the internal cycle algorithms described above has been analyzed only for separable demand functions: $d_i = d_i(s_i)$. In this case the conditions already discussed for the MSA-FA and MSA-CA algorithms for rigid demand equilibrium hold, with the further assumptions that the demand functions $d_i = d_i(s_i)$ are continuous, differentiable, non-decreasing monotone and bounded.

Among the internal cycle algorithms , the equivalent optimization problem (5.6.8) could be solved with the Frank-Wolfe algorithm for elastic demand symmetric deterministic equilibrium. However, with this approach it must be possible to express the inverse demand function, $\boldsymbol{Z}(\boldsymbol{d})$, between the minimum costs, $\boldsymbol{Z}$, and demand flows, $\boldsymbol{d}$, and this function must have a symmetric Jacobian. Both these conditions are difficult to meet in practice. In any case, the resulting algorithm requires some modifications of the DUN algorithm.

Hyper-network algorithms are based on the application of a rigid demand equilibrium assignment algorithm to an expanded network model which includes other links simulating choice behaviors on different dimensions as path choice behavior. From this point of view, hyper-network algorithms can be considered internal cycle algorithms. This approach can be applied only to some demand functions, and is briefly described below with reference to deterministic equilibrium. For the sake of simplicity, demand is assumed to be elastic only on the frequency dimension. Similar considerations can be made for elasticity on other choice dimensions, such as destination.

In particular, for each *od* pair, a fictitious path consisting of a single link is added to the network. For the demand conservation constraint, a flow equal to the excess demand flow, $h_k^0 = d_{od,max} - d_{od}$, is assigned to this path; this flow is equal to the potential demand flow not travelling (Fig.7.6.1). Let

$\boldsymbol{d}_{max}$ be the maximum demand flows vector;
$\boldsymbol{h}^0 = \boldsymbol{d}_{max} - \boldsymbol{d}$ be the vector of excess path flows;
$\boldsymbol{f}^0 = \boldsymbol{f}^0$ be the vector of excess link flows.

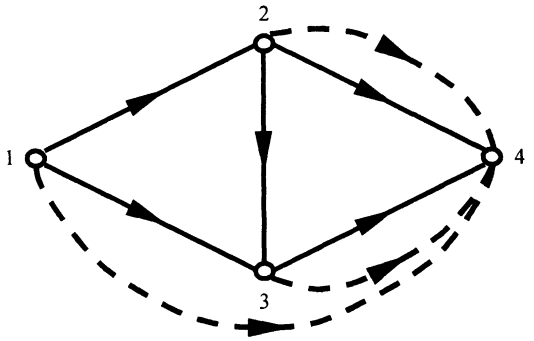


Fig. 7.6.1 Hyper-network approach.

A fictitious cost function can be associated to each new link, $c^0_{od} = c^0_{od}(\boldsymbol{f}^0)$. This function is obtained from the inverse demand function, relating minimum cost to demand flows as discussed in section 5.6.1.2:

$$Z_{od}(d) = Z_{od}(d_{max} - d_{max} + d) = Z_{od}(d_{max} - h^0) = g^0_{od}(\boldsymbol{h}^0) = c^0_{od}(\boldsymbol{f}^0)$$

It can easily be verified that the variational inequality model (5.4.3) for rigid demand deterministic equilibrium applied to this network is equivalent to the variational inequality model (5.6.6) for elastic demand deterministic equilibrium applied to the original network. Thus, the elastic demand DUE problem can be solved by applying a rigid demand DUE algorithm to the expanded network.

7.7. Applicative issues of assignment algorithms

In this chapter some algorithms for within-day static assignment have been described, with reference to models presented in Chapter 5 (whilst algorithms for within-day dynamic assignment presented in Chapter 6 are still at a reasearch stage).

Algorithms for (within-day static) traffic assignment are continually researched, exploring both the analysis of convergence conditions and implementation issues aimed at improving efficiency. It is therefore very difficult to give a complete picture of all algorithms that have been proposed to solve assignment problems, and the algorithms that have been described in this chapter are limited to those most commonly found in applications and easiest to implement. There are several variants to the basic specifications described, as well as more complex algorithms based on different approaches.

As noted throughout this chapter, even though stochastic assignment may seem more appealing from a theoretical point of view, solution algorithms are generally less efficient than those for deterministic assignment. On the other hand, there are very few algorithms for stochastic assignment, and, as noted in the bibliographic note, this topic is still worth of active research work.

Figure 7.7.1 presents, with a few modifications, the classification scheme of static assignment models described in Chapter 5, indicating for each class of models the main algorithms discussed in this chapter for rigid demand assignment.

TYPE OF NETWORK		PATH CHOICE MODEL	
		Deterministic	Stochastic Logit/Probit
Uncongested Network		All-or-Nothing	Dial / MonteCarlo
Congested Network	Symmetric User Equilibrium	Frank-Wolfe	MSA-FA
	Asymmetric User Equilibrium	Diagonalization	MSA-CA

Fig. 7.7.1 Classification of assignment algorithms

As a final comment, several commercial packages implementing at least rigid demand assignment models are available. Their use now spans over more than a decade and most of these packages can be considered reliable and robust.

Reference Notes

A general treatment of (static) assignment algorithms is given by the books by Sheffi (1985) and Patriksson (1994). Most of references quoted at the end of Chaper 5 also address solution algorithms. Some additional references are given below.

The literature proposes several algorithms for finding the shortest paths tree, which are useful for deterministic uncongested network assignment. A coprehensive treatment of algorithms for transportation networks and a comparison of their performances can be found in Gallo and Pallottino (1988), and in Ahuja, Magnanti and Orlin (1993).

Implementation of Stochastic Uncongested Network assignment algorithms is discussed in Sheffi (1985). For the Logit path choice model, the Dial algorithm described in section 7.3.1a is an original generalization of the algorithm described in the original work by Dial (1971); see also Van Vliet (1981). An adaptation of Dial's algorithm to C-Logit path choice model is described in Russo and Vitetta (1998). The Monte Carlo approach to stochastic uncongested network assignment was first proposed by Burrell (1968). Its application to Probit SUN assignment is described in Sheffi and Powell (1982). Maher and Hughes (1997)have proposed an approach to Probit SUN assignment based on Clark's approximation.

The MSA-FA algorithm for stochastic equilibrium is covered in Sheffi and Powell (1982), and its convergence is demonstrated in Powell and Sheffi (1982), as an optimization algorithm. Daganzo (1983) described the MSA-FA algorithm as a fixed-point algorithm, following Blum (1954), as well as the inverse cost function algorithm. The MSA-CA algorithm, and the internal cycle fixed-point algorithms for elastic demand assignment are covered in Cantarella (1997). External cycle MSA algorithms described in section 7.6 are an original contribution of this book. Other algorithms for the solution of Logit SUE symmetric models, based on the minimization model proposed by Fisk (1980), are described in Bell, Inaudi, Lam and Ploss (1993), Chen and Sule Alfa (1991), Damberg et al. (1996).

The adaptation of the Frank-Wolfe algorithm to the calculation of deterministic equilibrium flows is described in the original works of Le Blanc et al.(1975) and Nguyen (1976). As noted, many improvements to this algorithm have been proposed, such as the PARTAN, Florian and Spiess (1983), or other variations, Fukushima (1984), Lupi (1986). An interpretation of the Frank-Wolfe algorithm as a variational inequality algorithm is described in Van Vliet (1987). The diagonalization algorithm for non-separable cost functions is analyzed in Florian and Spiess (1982); other algorithms for non-separable cost functions are described in Nguyen and Dupuis (1984), Hearn, Lawphongpanich and Nguyen (1984).

The algorithm for computing shortest hyperpaths and its applications extension to DUE has been proposed by Nguyen and Pallottino (1988), see also Florian and Spiess (1989), and Wu, Florian, Marcotte (1994). The extension to stochastic assignment has been analyzed by Cantarella (1997) and Cantarella and Vitetta (2000). Algorithms for stochastic assignment with Logit hyperpath choice are described by Nguyen, Pallottino and Gendreau (1993). A comprehensive review of hyperpaths and related topics is in Gallo et al. (1993).

Notes

(1) The two problems are obviously equivalent since it is sufficient to change the directions of all the network links to obtain one problem from the other.

(2) Each centroid could be represented with a single node if the algorithms are appropriately modified to avoid a centroid node being crossed by a path.

(3) If some links have negative costs, loops with totally negative cost would lead to paths of minimum cost equal to minus infinite.

(4) A tree with root n, $T(n)$, is defined as a sub-graph of the whole graph with a single path correcting the root to each other mode. For further detail.

(5) This approach can be adopted for any random residual distribution.

(6) According to this assumption, a positive probability can be associated to a non- shortest path, given by the probability that the path is of maximum perceived utility (perceived shortest path). Given these considerations, SUN assignment is sometimes indicated as multi-path assignment in contrast with the all-or-nothing assignment for DUN, described in the next subsection.

(7) More generally, any set of relevant paths from the origin o can be adopted, as long as they form an acyclic graph, corresponding to a partial ordering of the nodes. In this case, in fact, a total ordering of the nodes can always be found, described by indices that take on the role of the distances $Z_{o,j}$ such that when a node is examined, all the preceding nodes have been examined.

(8) Using a simultaneous algorithm, given an origin (or a destination), independent of the tree structure, two sums for each link in the shortest paths tree are carried out, i.e. 2(n-1) additions if n is the number of nodes. Vice versa, using a sequential algorithm, the number of additions depends on the structure of the shortest paths tree. This number ranges between the number of the links of the tree, n-1, in the case that the paths within the tree do not overlap at all and the value $n_d(n-n_d-1) + n_d = n_d(n-n_d)$ (assuming $n>n_d$ with n_d the number of destinations) in the case of maximum overlapping.

(9) Some computational results suggest that the speed of convergence can be increased by reducing the step length by a factor $\beta \in]0, 1[$, $c^k = c^{k-1} + \beta/k\,(y^k - c^{k-1})$.

8 ESTIMATION OF TRAVEL DEMAND FLOWS

8.1. Introduction

Analysis and design of transportation systems require, respectively, the estimation of present demand and the forecasting of (hypothetical) future demand. These can be obtained by using different sources of information and statistical procedures.

To estimate the present demand, surveys can be conducted, typically by interviewing a sample of users; *direct estimates* of the demand can be derived using results from sampling theory.

Alternatively, the demand (present or future) can be estimated using models similar to those described in Chapter 4. *Model estimation* requires that the models are specified (i.e. the functional form and the variables are defined), calibrated (i.e. the unknown coefficients are estimated) and validated (i.e. the ability to reproduce the available data is tested). These operations can be performed on the basis of disaggregate information relative to a sample of individuals. The type of survey and the size of the sample are often different from those used for direct demand estimation. Once the models have been specified and calibrated, they can be applied to the present configuration of the activity and transportation systems to derive estimates of a present demand and/or to hypothetical configurations on the evolution of these systems (scenarios) to derive hypothetical forecasts of future demand.

Aggregate data can also be used for direct demand estimation and for the specification and calibration of demand models. Flows measured on network links are the most sophisticated form of aggregate data and can complement other disaggregate data and the relative estimation methods.

The different types of survey and estimation methodologies will be studied in the following sections of this chapter, as follows. Section 8.2 analyzes surveys and methods for direct demand estimation. Section 8.3 describes disaggregate estimations methods for the specification, calibration and validation of demand models based on traditional Revealed Preferences surveys. Section 8.4 describes some theoretical and operational aspects of Stated Preferences survey and calibration techniques, based on the information elicited from a sample of individuals in hypothetical scenarios. Sections 8.5 and 8.6 describe the methods using traffic counts to improve to estimate the present demand. Section 8.6 explores methods using traffic counts for aggregate *calibration* of demand models. Section

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

8.7 extends some of methods in previews section to deal with intra-periodal dynamic estimation. Finally, section 8.8 summarizes the methodologies for the estimation of the different components of transportation demand and discusses their fields of application. The topics listed are discussed for passenger transportation demand; extensions to freight demand are relatively straightforward.

8.2. Direct estimation of present demand

Transportation demand is the aggregation of individual trips made by the users of a given system under study. Full knowledge of the present demand would therefore require informations on the characteristics of the trips undertaken by all the users in the reference period (e.g. a typical day or part of it). Furthermore, as noted in Chapter 1, these informations should extend over several reference periods in order to compute average values. This census-like knowledge of transportation demand is neither practicable nor necessary. The associated economic and organizational costs, would make operation practically unfeasible in most cases. For these reasons, present transportation demand is typically estimated through sampling estimators, i.e. estimators based on information on a sample of system users.

In section 8.2.1, sampling surveys often used for direct demand estimation will be described; the estimators derived from sampling theory will be covered in section 8.2.2.

8.2.1. Sampling surveys

The basic idea of sampling techniques is to estimate relevant population variables on the basis of values observed in a relatively small group of individuals (sample) belonging to the population.

Several types of sample surveys can be used for direct estimation of transportation demand; these surveys, sometimes referred to as origin-destination surveys, may differ in their statistical characteristics and in the quality of information obtained. A comprehensive description of the various surveys is beyond the scope of this book; some typologies will be briefly described below as examples.

With *"while trip"* or *"on board"* surveys, a sample of users of one or more transportation modes are interviewed. The interviews can be conducted roadside for car drivers and their passengers, on board or at terminals (stations, airports, etc.) in the case of scheduled transportation services. The sample of users is obtained by randomly interviewing a predetermined fraction of the users of the mode chosen. In the case of "punctual" surveys (road sections, stations, etc.) this requires counting the total number of travelers passing the point (count of the universe) and interviewing a given number of them selected through a random mechanism. When on-board surveys are conducted to estimate exchange and crossing demand, they are also referred to as *cordon surveys*. In general, the information that can be gathered in these surveys is relatively "simple" since the interview has to be done in a short period of time and usually refers to the trip or journey under way.

With *household surveys*, a sample of families or persons living within the study area are interviewed. For families the sample is extracted randomly from the set of all resident families (simple random sample) or from the set of families living in each traffic zone (stratified random sample). The family members in the sample are interviewed about the trips taken in a given reference period. The same approach can be used for individuals rather than for families. The method of interviewing individuals in their homes usually is rather expensive but precise information are generally obtained because of the direct interaction between the interviewee and the interviewer. Household telephone surveys are becoming more and more popular, they have lower costs, although they usually yield less precise interviews.

There are several other types of sample surveys such as *destination surveys* in which travelers are interviewed at trip destinations (workplaces, schools, shops, etc.) and *(e)mail surveys* in which travelers are interviews by (e)mail. These surveys, though less costly than household surveys, may result in a potential bias of the estimates because of the systematic lack of information from some market segments.

The number of persons interviewed depends on the aims of the survey and the precision required for the estimates. Surveys aiming at direct estimation of the present demand usually require larger samples than those needed to calibrate demand models.

In applications different types of survey are employed to estimate different components of transportation demand; cordon surveys, for example, for exchange and cross trips and household surveys for internal trips.

Whatever the type of survey, the *statistical design* of a sampling survey for demand estimation consists of several standard phases:

- definition of the sampling unit (person, family, vehicle, etc.) and of the universe counting method (e.g. lists of residents or counts of passing vehicles);
- definition of the sampling strategy, i.e. the method for extracting the sample of individuals to be interviewed;
- definition of the estimator, i.e. the functions of the information obtained from the survey used to estimate the unknown quantities;
- definition of the number of units in (dimension of) the sample.

The *definition of the sampling unit* is largely influenced by practical matters such as the type of survey (household, on board, etc.) and the availability of information about the universe. For example, if the list of families living in a given area is available, but that of individuals is not, the sampling unit will be the family rather than the individual. In the case of on-board surveys the sampling unit will be the vehicle if the survey is carried out at the roadside or the passenger if the interviews are at the terminals.

For the *choice of sampling strategy*, almost all surveys make reference to *probabilistic sampling*, i.e. methods of sample extraction that define a priori the possible outcomes , assign a probability to each outcome, and extract randomly the

elements of the sample with this probability. In applications, the most commonly used probabilistic sampling strategies are:

- *simple random sampling*: all the elements of the population have an equal probability of belonging to the sample;
- *stratified random sampling*: the population is divided into non-overlapping, exhaustive groups (strata), subsequently a sample of elements is drawn from each stratum and each element of a stratum has an equal probability of belonging to the sample; elements in different strata may have different probabilities;
- *cluster sampling*: sampling units (e.g. people) are grouped in clusters (e.g. families or the passengers of a vehicle) and clusters are extracted randomly with a prefixed probability (simple random cluster sampling) or subdivided into strata and sampled with different probabilities (stratified random cluster sampling). The cluster sampling offers as a further possibility, the two-stage cluster sampling. In other words, a sample of clusters (e.g. a sample of families) is first extracted, subsequently a sample of individuals within each cluster is extracted. In this case, the probability of a unit belonging to the sample is the product of the probability of drawing the cluster to which it belongs and of the probability that the individual will then be extracted.

Choice of the estimator, i.e. the function of sample results, obviously depends on the variables to be estimated and on the sampling strategy adopted. In fact it can be demonstrated that an estimator that is statistically "efficient" for one strategy might not be for another.

The choice of the estimator and the definition of the sample size contain a stronger methodological content, discussed in the next subsection.

8.2.2. Sampling estimators

Present transportation demand can be estimated starting from the results of the sampling surveys described in the previous section. The problem of estimating Origin/Destination demand flows with certain characteristics (e.g. trip purpose and transport mode) and their main statistical properties for some sampling strategies will be addressed in the following.

Simple random sampling. In this case, a sample of n elements is drawn at random from a universe of N users. For example, in a household survey the sample of n families is obtained from the universe of the N families living in the study area. Let $d_{od}^{(1)}$ be the demand flow between origin o and destination d with given characteristics and n^{i}_{od} the number of these trips undertaken by the i-th element of the sample. Estimates of demand flows with given characteristics without distinguishing by origin-destination zones can be obtained in exactly the same way. Let n_{od} be the total of trips obtained from the sample. It obviously results:

$$n_{od} = \sum_{i=1,\dots n,} n^i_{od} \tag{8.2.1}$$

The sample estimate $\hat{d}_{od}$ of the demand flow for the whole universe can be obtained as follows:

$$\hat{d}_{od} = (N/n)n_{od} = (1/\alpha)n_{od} = N\bar{n}_{od} \tag{8.2.2}$$

where $\alpha=n/N$ is the *sampling rate* and $\bar{n}_{od} = n_{od}/n$ the average number of trips with the desired characteristics per element.

From sampling theory results, that an estimate of the variance[(2)] of $\hat{d}_{od}$ can be expressed as:

$$Var\left[\hat{d}_{od}\right] = N^2\hat{s}^2(1-\alpha)/n \tag{8.2.3}$$

where $\hat{s}^2$ is the sample estimate of the variance of the random variable n^i_{od} :

$$\hat{s}^2 = 1/(n-1)\sum_{i,=1\dots,n}\left(n^i_{od} - \bar{n}_{od}\right)^2 \tag{8.2.4}$$

In some surveys, a sample element (e.g. a car driver for cordon surveys) at the most undertakes one trip with the required characteristics (e.g. for a given purpose and/or in a given time band). In other surveys, the required information is whether the interviewee has a given characteristic (e.g. belongs to an income class or holds the driving license) or not. In both cases n^i_{od} is either zero or one and $\bar{n}_{od}$ is the *sampling estimate* of the percentage of travelers who have undertaken a trip of a certain type or have a given characteristic and will be indicated below by $\hat{P}_{od}$.

$$\hat{P}_{od} = \sum_{i,=1\dots,n} n^i_{od}/n \tag{8.2.5}$$

In this case the sampling estimate of the variance of n^i_{od} given by (7.2.4) can be expressed as the variance of a Bernoulli random variable:

$$\hat{s}^2 \cong \hat{P}_{od}(1-\hat{P}_{od}) \tag{8.2.6}$$

In fact, from (8.2.4), bearing in mind that in this case $n^i_{od}{}^2 \equiv n^i_{od,}$, it results:

$$\hat{s}^2 = \left[1/(n-1)\right]\sum_{i,=1\dots,n}\left(n^i_{o\,d}{}^2 + \bar{n}^2_{od} - 2n^i_{od}\bar{n}_{od}\right) \cong \hat{P}_{od} + \hat{P}^2_{od} - 2\hat{P}^2_{od} = \hat{P}_{od}\left(1-\hat{P}_{od}\right)$$

where the "almost equal" (≅) derives from assuming n equal to (n-1). In the case under study the estimate of the variance of $\hat{P}_{od}$ is given by:

$$Var\left[\hat{P}_{od}\right] = \hat{P}_{od}\ (1-\hat{P}_{od})(1-\alpha)/n$$

Stratified random sampling. In this case, the total population is divided into K groups of users, or strata; the generic stratum k has a population of N_k members and n_k elements are drawn at random from each stratum. This type of sampling is the most widely used in practical demand surveys. In cordon surveys, the strata includes users traveling through the different survey sections, while in household surveys the strata are often comprised of the families living in each zone (geographical stratification). In the first case the sample is "structurally" stratified because the users can be reached only in this way; in the second, the stratification is a choice made to guarantee a prefixed coverage of each zone.

If n^{ik}_{od} denotes the number of trips with the required characteristics undertaken by the i-th element in the sample of stratum k, an estimate of the total number of trips can be obtained as follows:

$$\hat{d}_{od} = N\sum_{k=1,\dots,K} w_k \sum_{i=1,\dots n_k} n^{ik}_{od} / n_k = N\sum_{k=1,\dots,K} w_k\ \bar{n}^{k}_{od} \qquad (8.2.7)$$

where n^{k}_{od} is the average number of trips observed in the k-th stratum, and $w_k = N_k/N$ is the weight of the stratum k with respect to the universe.

The variance of the stratified sampling estimate, $\hat{d}_{od}$, can be estimated as follows:

$$Var\left[\hat{d}_{od}\right] \approx N^2 \sum_{k=1,\dots,K} w_k^2 \hat{s}_k^2 (1-\alpha_k)/n_k \qquad (8.2.8)$$

where $\hat{s}_k^2$ is the sampling estimate of the variance of the variable n^{ik}_{od}:

$$\hat{s}_k^2 = 1/(n-1)\sum_{i=1,\dots,n_k} (n^{ik}_{od} - \bar{n}^{k}_{od})^2$$

and α_k is the sampling rate in the k-th stratum.

It can be shown that the sampling estimators of demand (8.2.2), (8.2.5) and (8.2.7) are unbiased and consistent estimators of the unknown demand if the interviews do not contain systematic distortions of the information provided (e.g. under-reporting of trips). The same can be said of the variance estimators (8.2.3) and (8.2.8). Variance estimates can be used to calculate the *confidence limits* of $\hat{d}_{od}$. If the sample is large enough to apply the central limit theorem, it can be assumed that

the estimator $\hat{d}_{od}$ follows a normal distribution. The upper and lower confidence limits of the estimate, $L^{S}_{1-\gamma}(d_{od})$ and $L^{I}_{1-\gamma}(d_{od})$, define the interval which, with probability (1-γ), includes the true value of d_{od}. On the assumption of a sufficiently large sample, these limits can be obtained as:

$$L^{S}_{1-\gamma}(d_{od}) = \hat{d}_{od} + z_{1-\gamma\,2} Var\left[\hat{d}_{od}\right]^{1/2}$$

and

$$L^{I}_{1-\gamma}(d_{od}) = \hat{d}_{od} + z_{\gamma/2} Var\left[\hat{d}_{od}\right]^{1/2}$$

where $z_{1-\gamma 2}$ and $z_{\gamma 2}$ are the 1-γ/2 and γ/2 percentiles of the normal standard variable. For γ=0.05, these percentiles are 1.96 and - 1.96 and the confidence limits are the extremes of the interval which with a probability of 0.95 contains the true value.

The ratio $IR(1-\gamma)$ between the width of the confidence interval and the value to be estimated is called *relative confidence interval* at (1 - γ) percent of the estimate $\hat{d}_{od}$:

$$IR(1-\gamma) = [L_{1-\gamma}{}^{S}(d_{od}) - L_{1-\gamma}{}^{I}(d_{od})]/d_{od} \qquad (8.2.9)$$

Expressions of the estimators and their variances for sampling strategies differing from the simple and stratified random sampling are more complex. However, the latter can still be used as first approximations. For the exact expressions of the estimators and of their variances in more complex sampling schemes, specialized texts in sampling theory should be consulted.

In principle, the sample size could be calculated according to the level of precision required by using expression (8.2.9) and substituting tentative values obtained from other studies for the variances $\hat{s}^2$ and $\hat{s}_k^2$ and for the variable d_{od}. For example, in the case of simple random sampling, if a relative $IR(1-\gamma)$ confidence interval of the estimate $\bar{n}_{od}$ is required at a given confidence level and the variation coefficient $\left(CV = s/\bar{n}_{od}\right)$ of the variable n^{i}_{od} is known, the sample dimension n can be obtained by combining expressions (8.2.2), (8.2.3) and (8.2.9) as follows:

$$n \approx 4\frac{CV^2 z^2_{1-\gamma/2}(1-\alpha)}{IR(1-\gamma)^2} \qquad (8.2.10)$$

A similar expression can be obtained for a given relative confidence interval of the O-D demand flows $\hat{d}_{od}$.

In practice, the theoretical computation of the sample size is rarely possible because several parameters are estimated from the same survey. Furthermore, the

sample size required for sufficiently precise estimates of some parameters, and especially of the single elements of an O-D matrix, would be too large to be feasible. The usual practice is to choose a sample size used with other "successful" surveys, verifying that some aggregate estimates (e.g. the level of demand or the number of trips in each zone for each purpose) have a satisfactory minimum precision.

As an example, Fig. 8.2.1 shows the sampling rate[3] for urban household origin-destination surveys recommended by the U.S. Bureau of Public Roads as a function of the resident population.

Finally, it should be noted that the use of models as estimators of present demand is becoming more and more widespread (see section 8.8). This is due to the low level of precision that can be achieved by direct estimates and, on the other hand, to the effectiveness of specification and calibration techniques of demand models.

RESIDENT POPULATION		SAMPLING RATE	
		Recommended	Minimum
Less than 50.000		0.200	0.100
50.000	150.000	0.125	0.050
150.000	300.000	0.100	0.030
300.000	500.000	0.067	0.020
500.000	1.000.000	0.050	0.015
More than 1.000.000		0.040	0.010

Fig. 8.2.1 Sampling rates for household surveys in relation to resident population (BPR-USA)

8.3. Disaggregate estimation of demand models

Estimation of transportation demand by means of mathematical models, whether they are applied to the present situation or to hypothetical scenarios, requires the specification, calibration and validation of these models. In other words, it is necessary to define the functional form and the variables (attributes) included in the model, to estimate the coefficients or parameters, and to verify the "statistical quality" of the model. A good demand model is usually the outcome of a process of trial and error in which the specification-calibration-validation cycle is repeated several times until a satisfactory result is obtained. In this process the modeler's judgment and experience play a central role.

These operations, which will be called synthetically *model estimation*, can be performed starting from information on the travel behavior of a sample of users. This approach is called disaggregate estimation[3] of demand models. The surveys used to gather elementary information might belong to two different classes: surveys relative to the actual travel behavior in a real context (*Revealed Preferences* or *RP* surveys) or surveys relative to the hypothetical travel behaviors in fictitious scenarios (*Stated Preference* or *SP* surveys). The traditional method of revealed preferences is based on surveys analogous to those described in section 8.2.1. These surveys provide information on users' choice relevant for the model to be calibrated (e.g. the transportation mode chosen for the calibration of a modal choice model).

Survey design therefore consists of the definition of the sample size, the questionnaire and the sampling strategy. Stated preferences (SP) surveys differ in that they are conceptually equivalent to a laboratory experiment designed with a larger number of "degrees of freedom". Given the complexity of the subject, SP survey design and their use for the calibration of demand models will be covered in the next section. What follows will therefore consider the specification, calibration and validation of demand models with reference to a generic RP survey.

Independent of their interpretation (behavioral or descriptive) and functional form, demand models can be seen as mathematical relationships, which give the probability that a generic user i, chooses a particular travel option among those available. Thus a mode choice model $p[m/ods]$ expresses the probability that a user, randomly selected from those who undertake a trip for the purposes s between zones o and d pair, uses mode m. This section will addresses the problem of building demand models, or systems of models, making reference to a generic model expressing the probability $p^i[j]$ that a user i, chooses the travel option j among those available.

Section 8.3.1 will discuss some general considerations on model specification. Section 8.3.2 will cover calibration methods, and finally section 8.3.3 will describe some validation methods.

8.3.1. Model specification

The specification of a demand model can be defined as the complete identification of its mathematical structure, i.e. the definition of its functional form and of the explanatory variables (attributes) used.

The choice of the model's functional form, e.g. Multinomial Logit or Hierarchical Logit, depends on many factors such as its computational tractability, the results obtained in similar cases, the a priori expectations on the correlation of random residuals. In general the assumptions made can be tested a posteriori on the basis of the statistical tests described in section 8.3.3.

The choice of the explanatory variables clearly depends on the specific type of model. However, there are some rules that should be observed to avoid problems in the calibration phase. In general variables that are *collinear,* i.e. linearly dependent on each other, should be avoided. In fact, if the systematic utility function is linear with respect to collinear attributes, there are infinite combinations of their coefficients giving equal values of systematic utilities and of choice probabilities. This circumstance makes it impossible to estimate (identify) separately the related coefficients in the phase of model calibration. *Alternative Specific Attributes* in additive random utility models are a typical example of linearly dependent, or multi-collinear, variables and, as was seen in Chapter 3, should be introduced at the most for all the alternatives but one. Socio-economic attributes can also be a source of multi-collinearity. A socio-economic characteristic, such as income or car ownership, is constant for all the alternatives and can therefore be introduced, at the most, in the systematic utility function of all the alternatives except one. In any case socio-economic variables should not be employed as alternative specific attributes.

For example, two "high income" dummy variables should not be introduced in the systematic utilities of the alternatives car and taxi with different coefficients. A third type of collinearity might be introduced when one attribute is obtained from another; this would be the case if the travel time was deduced from the distance by assuming a constant speed, travel time and distance should not be included in the model specification as two distinct variables.

8.3.2. Model calibration

Random utility models can be seen as mathematical relationships expressing the probability $p^i[j](\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\theta})$ that individual i chooses alternative j as a function of the vector $\boldsymbol{X}$ of attributes for all the available alternatives and of the vectors of parameters relative to the systematic utility, $\boldsymbol{\beta}$, and to the joint probability function of the random residuals, $\boldsymbol{\theta}$. Choice probabilities depend on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ through systematic utility functions, usually specified as linear combinations of the attributes $\boldsymbol{X}$ (or their transformations) with coefficients given by the parameters $\boldsymbol{\beta}$:

$$V_j(\boldsymbol{X}^i_j) = \sum_z \beta_z X^i_{zj} = \boldsymbol{\beta}^T \boldsymbol{X}^i_j \tag{8.3.1}$$

Structural parameters θ include all parameters related to the probability distribution function of random residuals. Thus, in the case of the Multinomial Logit models, θ is the only parameter θ of the Gumble random variables. In the Hierarchical Logit, $\boldsymbol{\theta}$ consists of parameters θ_o and θ_r associated to structural nodes. In the Probit model, $\boldsymbol{\theta}$ consists of all the elements of the variance-covariance matrix and so on.

Calibrating the model requires the estimation of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from the choices made by a sample of users. It should be observed that in general not all the coefficients can be identified, i.e. they can be estimated separately. We shall return to this point in greater detail later with reference to specific examples.

The Maximum Likelihood Method. Maximum Likelihood (*ML*) is the method most widely used for estimating model parameters. In Maximum Likelihood estimation the values of the unknown parameters are obtained by maximizing the probability of observing the choices made by a sample of users. The probability of observing these choices, i.e. the likelihood of the sample, depends (in addition to the choice model adopted) on the sampling strategy adopted. The cases of simple and stratified random sampling will be considered in the following.

In the simplest case of *simple random sampling* of n users, the observations are statistically independent and the probability of observed choices is the product of the probabilities that each user i chooses $j(i)$, i.e. the alternative actually chosen by him/her. The probabilities $p^i[j(i)](\boldsymbol{X}^i; \boldsymbol{\beta}, \boldsymbol{\theta})$ are computed by the model and therefore depend on the coefficients vectors. Thus, the probability L of observing the whole sample is a function of the unknown parameters:

$$L(\beta,\theta) = \Pi_{i=1,\ldots,n}\, p^i[j(i)](X^i;\beta,\theta) \qquad (8.3.2)$$

The Maximum Likelihood estimate $[\beta,\theta]_{ML}$ of the vectors of parameters β and θ is obtained by maximizing (8.3.2) or, more conveniently, its natural logarithm (log-likelihood function):

$$[\beta,\theta]_{ML} = argmax\, lnL(\beta,\theta) = argmax \sum_{i=1,\ldots,n} ln\, p^i\left[j(i)\right]\left(X^i,\beta,\theta\right) \qquad (8.3.3)$$

Fig. 8.3.1 shows an elementary example of the Maximum Likelihood estimation of a single parameter.

In the calibration of some models, the n users may be grouped in sets of n_i users, all choosing the same alternative and having the same attributes. A typical example is an aggregate distribution model in which the users travelling between the same O-D pair possess the same attributes, namely the trip costs between zone pairs and attraction variables of each destination. In this case the likelihood function and its logarithm can be expressed as:

$$L(\beta,\theta) = \Pi_i\, p^i\left[j(i)\right]^{n_i}(X^i,\beta;\theta)$$
$$ln\, L(\beta,\theta) = \sum_i n_i\, ln\, p^i\left[j(i)\right](X^i;\beta,\theta)$$

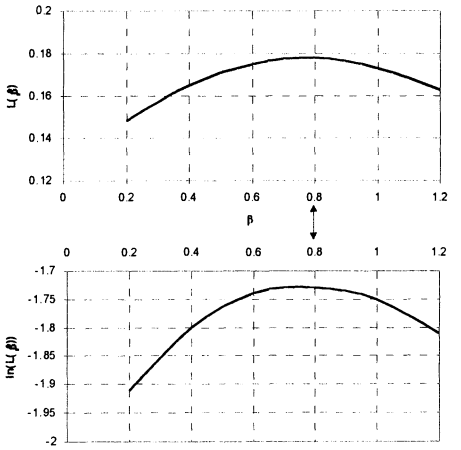
In *stratified random sampling*, n_h users are sampled randomly from the N_h members of each stratum ($h = 1,\ldots,H$) with a sampling rate $\alpha_h = n_h/N_h$. The probability of observing sample choices and therefore the likelihood function, depends on the method used to identify the strata.

If the population is stratified using, either directly or indirectly, the attributes $\boldsymbol{X}$ but not the choices to be modeled, the strategy is known as *exogenous stratified sampling*. Typical examples are geographical stratification (level of service attributes depend on the zone or the zone pair on which the stratification is carried out) and/or income stratification.

$$n=3 \quad j=A,B \quad p[A]=\frac{\exp(-\beta C_A)}{\exp(-\beta C_A)+\exp(-\beta C_B)}$$

user	$j(i)$	C_A^i	C_B^i
1	A	3	5
2	A	2	1
3	B	4	3

$$L(\beta)=\frac{\exp(-3\cdot\beta)}{\exp(-3\cdot\beta)+\exp(-5\cdot\beta)}\cdot\frac{\exp(-2\cdot\beta)}{\exp(-2\cdot\beta)+\exp(-1\cdot\beta)}\cdot\frac{\exp(-3\cdot\beta)}{\exp(-3\cdot\beta)+\exp(-4\cdot\beta)}$$



β	$p^1[A]$	$p^2[A]$	$p^3[B]$	$L(\beta)$	$ln(L(\beta))$
0.20	0.60	0.45	0.55	0.148	-1.91
0.40	0.69	0.40	0.60	0.165	-1.80
0.60	0.77	0.35	0.65	0.175	-1.74
0.80	**0.83**	**0.31**	**0.69**	**0.178**	**-1.73**
1.00	0.88	0.27	0.73	0.173	-1.75
1.20	0.92	0.23	0.77	0.163	-1.81

Fig. 8.3.1 Maximum Likelihood estimation of a single parameter.

For samples obtained through exogenous stratified sampling it can be demonstrated that the log-likelihood function is:

$$ln\, L(\beta,\theta)=\sum_{h=1,\ldots,H}\sum_{i=1,\ldots,n_h} ln\, p^i[j(i)](\mathbf{X}^i;\beta,\theta)+const. \qquad (8.3.4)$$

which, apart from a constant term, coincides with the function (8.3.3) obtained for a simple random sample with size $n=\Sigma_{h=1,\ldots,H} n_h$

If the stratification is based on the choices made by the users, the sampling strategy is knows as *choice-based stratified sampling*. This is the case, for example, if the sample used to calibrate a mode choice model is obtained by randomly selecting a sample of users of each transport mode; the population of each stratum is comprised of all users choosing each mode. Specify the log-likelihood function in closed form exactly is rather complex for this sampling strategy. As an approximation, the Maximum Likelihood estimator with *exogenous weights* can be adopted; in this case the function $ln\ L(\beta,\theta)$ is expressed as:

$$ln\ L(\beta,\theta) = \sum_{h=1,\ldots,H} \left(\frac{w_h}{\alpha_h} \right) \sum_{i=1,\ldots,n_h} ln\ p^i[j(i)](X_j^i;\beta,\theta) \qquad (8.3.5)$$

which, apart from the weights $w_h\alpha_h$, coincides with (8.3.4) and therefore with (8.3.3). In the previous expression each observation is weighed according to the ratio between the weight of the stratum w_k (ratio between the population of the stratum and the total population) and the sampling fraction of the stratum itself. Thus the observations are corrected so that the elements belonging to under-sampled strata receive greater weights than those belonging to over-sampled strata. Furthermore, the weighted log-likelihood function (8.3.5) coincides with (8.3.4) if the sampling fraction of each stratum is proportional to the weight of the stratum ($w_h\alpha_h=const.$). To apply the Maximum Likelihood estimator with exogenous weights to a choice-based stratified sample, it is therefore necessary to have an estimate of the weight of each stratum, i.e. of the fraction of the total population choosing each alternative. This information can be obtained from official statistics, or estimated from another simple random sample smaller in size and/or with less detailed questionnaires.

From the statistical point of view, under some assumptions, Maximum Likelihood estimators have many asymptotic properties such as consistency, efficiency and normality, independent of the model expressing the probabilities $p^i[j]$. Furthermore, it is possible to obtain approximate estimates of the variances and covariances of the components of β_{ML}, since its dispersion matrix Σ is asymptotically equal to minus the inverse of the log-likelihood function Hessian calculated at point $(\beta,\theta)_{ML}$:

$$\sum_{\beta,\theta} = -\left[\frac{\partial^2 ln\ L(\beta,\theta)}{\partial(\beta,\theta)\,\delta(\beta,\theta)^T} \right]^{-1}_{(\beta,\theta)_{ML}} \qquad (8.3.6)$$

If the sample is sufficiently large, expression (8.3.6) can be used to estimate variances and confidence limits for the coefficients.

From the algorithmic point of view, Maximum Likelihood estimation requires the solution of an unconstrained maximization problem, expressed by (8.3.3) or (8.3.5). This problem can be solved by applying a gradient algorithm of the type described in Appendix A. The gradient of the objective function can be calculated analytically or numerically depending on the functional form of the model $p^i[j(i)]$ to be calibrated.

Maximum Likelihood Estimators for some random utility models. The explicit formulation of the functions $ln\ L(\beta,\theta)$ in expressions (8.3.3), (8.3.4) and (8.3.5), the possibility of estimating the coefficients, as well as the properties of the unconstrained optimization problem depend on the type of model used. The cases of the Multinomial Logit and the Hierarchical Logit models can be treated analytically and are described below.

If the probabilities $p^i[j](\boldsymbol{X}^i;\beta,\theta)$ are obtained with a *Multinomial Logit* model with a systematic utility linear in the coefficients β_k, the objective function (8.3.3) can be expressed analytically:

$$ln\ L(\beta,\theta) = \sum_{i=1,\dots,n}\left[\sum_{k=1,\dots,K}\beta_k X^i_{kj(i)}/\theta - ln\sum_{j\in I_i} exp\left(\sum_{k=1,\dots,K}\beta_k X^i_{kj}/\theta\right)\right]$$

or in vector form:

$$lnL(\beta,\theta) = \sum_{i=1,\dots,n}\left[\beta^T \boldsymbol{X}^i{}_{j(i)}/\theta - ln\sum_{j\in I_i} exp(\beta^T \boldsymbol{X}^i_j/\theta)\right] \qquad (8.3.7)$$

In this case the parameters to be estimated are the N_β coefficients β_k, plus a single parameter θ. As previously noted, not all parameters can be estimated separately since the values of the log-likelihood function (8.3.7) do not depend on the N_β+1 single parameters but on N_β ratios β_k/θ. It can be immediately verified, in fact, that a vector $[\beta_1, \beta_2 \dots \theta]$, and a vector $[\alpha\beta_1, \alpha\beta_2,\dots,\alpha\theta]$, give the same value of the function (8.3.7). Thus, it would be impossible to estimate β_k and θ separately, since there are infinite combinations of them giving the same choice probabilities and therefore the same value of the log-likelihood function. If the ratio β_k/θ is denoted by β_k', the vector β' is:

$$\beta' = \beta/\theta = [\beta_1/\theta, \beta_2/\theta,\dots]$$

and expression (8.3.7) becomes[4]:

$$ln\ L(\beta') = \sum_{i=1..n}\left[\beta'^T \boldsymbol{X}^i{}_{j(i)} - ln\sum_{j\in I_i} exp(\beta'^T \boldsymbol{X}^i_j)\right] \qquad (8.3.8)$$

The first partial derivatives of (8.3.8) with respect to the generic parameter β_k' can be used to compute the gradient of the objective function and can be expressed in closed form:

$$\frac{\partial \ln L(\beta')}{\partial \beta_k'} = \sum_{i=1\ldots n} \left[X^i_{kj(i)} - \sum_{j \in I_i} X^i_{kj} \frac{exp\left(\beta'^T X^i_j\right)}{\sum_{h \in I_i} exp\left(\beta'^T X^i_h\right)} \right]$$

or in a more compact notation:

$$\frac{\partial \ln L(\beta')}{\partial \beta_k'} = \sum_{i=1\ldots n} \left[X^i_{kj(i)} - \sum_{j \in I_i} X^i_{kj} p^i[j](X^i, \beta') \right] \tag{8.3.9}$$

Also the partial second order derivatives of $\ln L(\beta')$ can be expressed in closed form.

$$\frac{\partial^2 \ln L(\beta')}{\partial \beta_k' \partial \beta_l'} = -\sum_{i=1\ldots n} \sum_{j \in I_i} p^i[j](\beta') \cdot \left(X^i_{jk} - \sum_{h \in I_i} X^i_{hk} p^i[h] \right) \cdot \left(X^i_{jl} - \sum_{h \in I_i} X^i_{hl} p^i[h] \right) \tag{8.3.10}$$

These derivatives can be used in some algorithms to solve the optimization problem (8.3.3) and to obtain a sample estimate of the variance-covariance matrix Σ_{ML} of the estimator β'_{ML} given by (8.3.6).

Under rather general assumptions, it can be shown that the Hessian matrix of the objective function (8.3.8), whose components are given by the second derivatives (8.3.10), is definite negative and thus the function $lnL(\beta')$ is strictly concave. Therefore there is a unique vector β'_{ML} maximizing the function $ln\ L(\beta')$ and the algorithms described in Appendix A converge to this value.

These results can be extended to the case of functions $lnL(\beta')$ given by (8.3.4) and (8.3.5) for stratified samples.

In the case of Hierarchical Logit models, choice probabilities depend on the structure of the choice tree. For the sake of simplicity, reference will be made to the example in Fig. 8.3.2 in which the parameters θ and δ relative to structural nodes are indicated. The results can be extended to any choice tree structure.

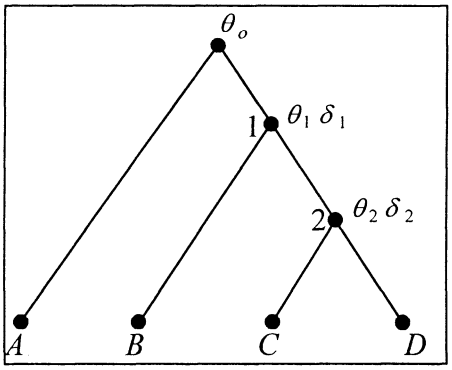


Fig. 8.3.2 Choice tree structure for a Nested Hierarchical model.

In this case we have:

$$p[A]=\frac{e^{V_A/\theta_o}}{e^{V_A/\theta_o}+e^{\theta_1 Y_1/\theta_o}}=\frac{e^{V_A/\theta_o}}{e^{V_A/\theta_o}+\left[e^{V_B/\theta_1}+\left(e^{V_C/\theta_2}+e^{V_D/\theta_2}\right)^{\theta_2/\theta_1}\right]^{\theta_1/\theta_o}}$$

Substituting the expressions relative to the systematic utilities, it results:

$$p[A]=\frac{e^{\sum_k \beta_k X_{kA}/\theta_o}}{e^{\sum_k \beta_k X_{kA}/\theta_o}+\left[e^{\sum \beta_k X_{kB}/\theta_1}+\left(e^{\sum_k \beta_k X_{kC}/\theta_2}+e^{\sum_k \beta_k X_{kD}/\theta_2}\right)^{\theta_2/\theta_1}\right]^{\theta_1/\theta_o}} \tag{8.3.11}$$

Choice probabilities, and the log-likelihood function, in addition to the N_β coefficients β_k, depend on N_θ parameters θ_r, one for each intermediate node plus one (θ_o) for the root. It can also be observed that the structural coefficients always appear in (8.3.11) as ratios. Each coefficient β_{kj} in the systematic utility of an alternative j is divided by the parameter $\theta_{a(j)}$ relative to the parent node of j, while each parameter θ_r relative to an intermediate node r is divided by the parameter $\theta_{a(r)}$ relative to the parent node of r which may be an intermediate node or the root.

For Hierarchical Logit models, the $N_\beta+N_\theta-1$ ratios, rather than the single $N_\beta+N_\theta$ parameters, can be calibrated. In fact, it can be verified immediately that a vector $[\beta_1, \beta_2,...,\beta_{N_\beta},\theta_1,\theta_2,...,\theta_{N_\theta}]$, and a vector $[\alpha\beta_1,\alpha\beta_2,...,\alpha\beta_{N_\beta},\alpha\theta_1,\alpha\theta_2,...,\alpha\theta_{N_\theta}]$ substituted in expression (8.3.11) give the same value of $p[A]$. All the parameters can therefore be identified but one. The parameters usually identified are the ratios $\beta'_{kj}=\beta_{kj}/\theta_{a(j)}$ and $\delta_r=\theta_r/\theta_{a(r)}$[5].

From the previous expressions it is also deduced that the coefficients β_k of a generic attribute appearing in the utilities of alternatives belonging to different nests, for example β_{kA} and β_{kC}, must satisfy a consistency relationship:

$$\beta_{kA} = \beta_{kC} \Rightarrow \beta'_{kA} = \beta'_{kC}\,\delta_1\,\delta_2$$

For these considerations, if the vector of the ratios $\beta_{kj}/\theta_{a(j)}$ is denoted by β' and the vector of the ratios $\theta_r/\theta_{a(r)}$ by δ, the log-likelihood function becomes $lnL(\beta',\delta)$. It can be shown that this function is concave with respect to the vector β', for a given δ, while not concave with respect to the vector δ. Fig. 8.3.3 shows the graph of the objective function $lnL(\beta',\delta)$ for a simple Hierarchical Logit model as a function of a single parameter δ where the vector β' is equal to the (unique) value maximizing the log-likelihood function for the value of δ in abscissa. The figure shows the non-concavity of the function and two local maxima. For this reason, the problem (8.3.3) is solved sometimes by using heuristic algorithms which maximize the log likelihood function with respect to the vector β' for a set of fixed values of δ and subsequently search within the limited set of trial vectors δ (e.g. grid search). Other algorithms solve the problem (8.3.3) directly with appropriate definition of the ascent direction.

Another possibility for the calibration of Hierarchical Logit models is the sequential estimation of the parameters of Multinomial Logit models corresponding to each node of the choice tree associated with the decision process. The calibration process is started from the intermediate nodes, which include only elemental alternatives. Parameters calibrated at one stage are kept fixed in the following stages. This type of calibration is known as *sequential* or *partial information estimation*, since for each calibration the only information used is relative to users who have chosen elemental alternatives (leaves) of the tree and/or compound alternatives (structural nodes) connected to the intermediate node under study. There are, however, both theoretical and practical problems connected with partial information Maximum Likelihood estimation. From the theoretical point of view, the method is sub-optimal, i.e. it can produce a value of the objective function which is lower than the global maximum. Furthermore, the values of the objective function are sometimes lower than those obtained calibrating the Multinomial Logit model with equal systematic utilities. This is clearly a contradiction since the latter is a special case of the Hierarchical Logit model with all δ's equal to one. From the practical point of view, in sequential estimation it is very difficult to estimate the coefficients of generic attributes with the same coefficient β_k, if introduced in the systematic utilities of alternatives belonging to different groups. In fact, each group is calibrated separately and it is not easy to impose equality constraints between parameters common to two or more groups. For these reasons the sequential estimation is not to be recommended.

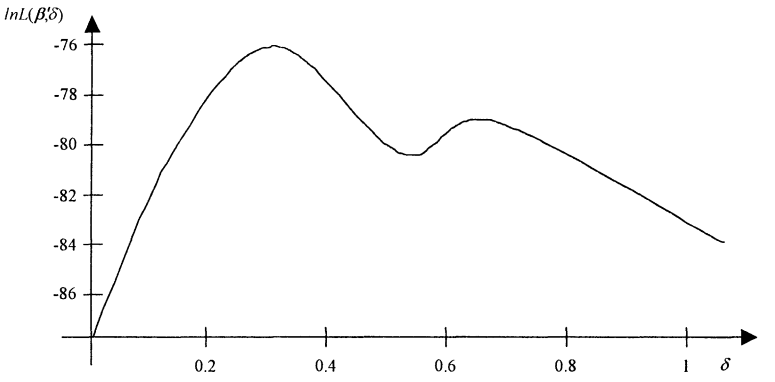


Fig. 8.3.3 Log-Likelihood for a Hierarchical Logit model as a function of the parameter δ.

8.3.3. Model validation

Once a demand model has been specified and calibrated, it must be validated. In this phase the reasonableness and the significance of estimated coefficients are verified, as well as the model's capability to reproduce the choices made by a sample of users. In addition, the assumptions underlying the functional form assumed by the model are tested. All of these activities can be completed with appropriate tests of hypotheses for a sample of users.

Informal tests on coefficients. These tests are based on the expectations on the signs of the coefficients calibrated and on their reciprocal relationships.

Wrong signs of the coefficients are likely indicators of errors in the attribute database in survey results, or of the model mis-specification. For example, in a road path choice model, it may happen that paths including toll motorway sections are chosen, even though they have approximately the same average travel time and are more expensive. If the model specification does not account for the greater driving comfort, the calibration procedure may result in a positive cost coefficient to increase the systematic utility, and therefore the choice probability, of motorway alternatives. A different specification of the model introducing an attribute equal to the length of the motorway section of each path should adjust the cost coefficient to the expected negative value.

Other checks can be conducted on the ratios between the coefficients of different attributes. As stated in Chapter 4 (see equation (4.2.35)), the ratio between time and monetary cost coefficients can be interpreted as *Value of Time* (VOT) and can be compared with the results of other calibrations and the expectations about the users' willingness to pay. The parameters of attributes corresponding to different components of travel time (e.g. waiting and on board time) should have increasing

absolute values for less appreciated components and so on. In general, the results reported in the scientific and technical literature are very helpful for these analyses.

As an example, consider the modal choice model described in Fig. 8.3.4; the times and cost coefficients are negative, while the availability coefficients (car, motorcycle and bicycle) are positive. Furthermore, the perceived time value is about 5 €/hr. It can also be seen that the disutility associated with time on board and so on.

$$
\begin{aligned}
V_{Car} &= \beta_T \cdot T_a + \beta_{CA} \cdot CA + \beta_{HF} \cdot HF + \beta_{CAR} \cdot CAR \\
V_{motorbike} &= \beta_T \cdot T_m + \beta_{MAN} \cdot MAN + \beta_{21-35} \cdot 21-35 \\
V_{bus} &= \beta_T \cdot T_b + \beta_{Taccb<10} \cdot T_{accb<10} \\
V_{walking} &= \beta_{Twlk} \cdot T_{wlk} + \beta_{WLK} \cdot WLK
\end{aligned}
$$

T_c, T_m, T_b = travel times of the modes "car", "motorcycle", "bus";
T_{wlk} = walking travel time;
$T_{acc<10}$ = dummy variable = 1 if access time to bus is less than ten minutes, 0 otherwise;

CA = car availability (no. Cars/no. licenses in the household);
MAN = dummy variable = 1 if the user is male, 0 otherwise;
HF = dummy variable = 1 if the user is head of family, 0 otherwise;
$21\text{-}35$ = dummy variable = 1 if the user is aged between 21 and 35, 0 otherwise
CAR, WLK = Alternative Specific Attributes (ASA);

Coefficients	β_T	β_{Twlk}	$\beta_{Tacccb<10}$	β_{CA}	β_{HF}	β_{MAN}	β_{21-35}	β_{CAR}	β_{WLK}
Estimate	-0.748	-4.560	1.247	1.758	0.452	0.990	1.684	1.411	3.929
Std.dev.	0.338	0.431	0.472	0.384	0.225	0.532	0.466	0.560	0.548
T	-2.213	-10.59	2.642	4.573	2.012	1.962	3.616	2.519	7.168

Test	H	Test statistic	95th percentile
t student	$\beta_t = \beta_{Twlk}$	7.53	1.96
$LR(\mathbf{0})$	$\beta = \mathbf{0}$	588.01	16.92
$LR(\beta_{ASA})$	$\beta = \beta_{ASA}$	285.83	14.06

Goodness of fit test	
ρ^2	0.424
$\bar{\rho}^2$	0.411

Fig. 8.3.4 Parameters and statistics for a modal choice Logit model.

Formal tests on coefficients. Under the assumption of sufficiently large samples, different assumptions on Maximum Likelihood estimates $\beta^{ML(6)}$ can be tested using asymptotic results.

t-student tests on particular coefficients

These tests check the null hypothesis that a coefficient β_k is equal to zero and the estimate β_k^{ML} differs from zero for sampling errors ($H_0 : \beta_k = 0$) through the statistic:

$$t = \frac{\beta_k^{ML}}{Var\left[\beta_k^{ML}\right]^{1/2}} \tag{8.3.12}$$

Alternatively, the t-student statistic can be used to test that two coefficients β_k and β_j are equal ($H_0 : \beta_k = \beta_j$):

$$t = \frac{\beta_k^{ML} - \beta_j^{ML}}{\left(Var\left[\beta_k^{ML}\right] + Var\left[\beta_j^{ML}\right] - 2\,cov\left[\beta_j^{ML}\beta_k^{ML}\right]\right)^{1/2}}$$

In both cases, under the null hypothesis the statistic t is distributed according to a t-student variable with a number of degrees of freedom equal to the size of the sample minus the number of coefficients estimated. Given the typical sample size it is usually assumed that the t statistic is distributed as a normal standard variable, $N(0,1)$, which is the limit distribution of the t-student variable. Sample estimates of variances and covariances can be computed through expression (8.3.6). It is well known that the null hypothesis is rejected with a probability α of making a Type I error (e.g. rejecting a true assumption) if the value of the t statistic is external to the extremes of the interval $\left(z_{\alpha/2}, z_{1-\alpha/2}\right)$ which for α=0.95 are equal to ± 1.96. The values of the t-student statistics (8.3.12) for the coefficients of the model reported in Fig. 8.3.4 show that all the estimates of the coefficients are significantly different from zero with α=0.95. The reader can check the significance of the coefficients of the different models described in Chapter 4.

Chi-square tests on vectors of coefficients

To test for the null hypothesis that the true vector β of the coefficients or one of its sub-vector is equal to a given vector β^*, ($H_0 : \beta = \beta^*$), the following statistic can be used:

$$chi^2(\beta^*) = (\beta^{ML} - \beta^*)^T \sum_{\beta}^{-1} (\beta^{ML} - \beta^*) \tag{8.3.13}$$

If the null hypothesis is true, chi^2 is asymptotically distributed as a chi-square variable with a number of degrees of freedom equal to the number of components of β.

Note that expressions (8.3.12) and (8.3.13) can be used to obtain the confidence interval of a single coefficient β_k and the confidence region of a vector of coefficients.

Likelihood Ratio tests on vectors of coefficients. The Likelihood Ratio test is similar to the previous one and tests the null hypothesis that the vector β, or one of its sub-vectors, is equal to a vector β^*. The vector β^* may be defined implicitly imposing some constraints on β. Constraints can be synthetically represented by a feasibility set B defined by them ($\beta \in B$). Both in the implicit and the explicit case, β^* can be seen as the vector maximizing the log-likelihood function under the constraints:

$$\beta^* = \underset{\beta \in B}{arg\,max}\ ln\, L(\beta)$$

For instance, one can test that β is null, $\beta^*=\mathbf{0}$ or that only some of its components are null; in the latter case the other components of β^* will be estimated by solving the constrained maximization problem.

The null hypothesis $H_o:\beta=\beta^*$ can be tested using the *Likelihood Ratio* statistic *LR*:

$$LR\,(\beta^*)= -2\,[ln\, L(\beta^*) - ln\, L\,(\beta^{ML})] \qquad (8.3.14)$$

which, on the null hypothesis, is asymptotically distributed according to a chi-square variable with a number of degrees of freedom equal to the number of constraints imposed in estimating β^*.

The *LR* statistic is always greater than zero since the unconstrained maximum $lnL(\beta^{ML})$ of the function $lnL(\beta)$ is not smaller than the constrained maximum of the same function, $ln\, L(\beta^*)$. Note that the *LR* test is equivalent, but not equal from the numerical point of view, to the chi-square test described above when the constraints completely identify the vector β^*. For example, in the case $\beta^*=\mathbf{0}$ it yields:

$$LR(\mathbf{0})= -2[ln\, L(\mathbf{0}) - ln\, L\,(\beta^{ML})] \qquad (8.3.15)$$

The null hypothesis $\beta^*=\mathbf{0}$ corresponds to assuming a "true" model with all coefficients equal to zero and therefore with equiprobable alternatives ($V_j = 0\ \forall\ j \Rightarrow p[j] = 1/J$). This hypothesis is the less likely the larger the difference between the probability of observing the users' choices with the calibrated model ($lnL(\beta^{ML})$ and that with a zero coefficients model ($lnL(\mathbf{0})$). Under the null hypothesis the statistic $LR(\mathbf{0})$ will be distributed as a chi-square variable with a number of degrees of freedom equal to N_β

A more challenging specification of the test is obtained by comparing the calibrated model with a model whose only parameters are the Alternative Specific Attributes β_{ASA}. The vector $\beta^*=\beta^{ML}{}_{ASA}$ is obtained by maximizing the log-likelihood function $lnL(\beta)$ with the constraints that all the other coefficients are equal to zero: the number of ASA and their coefficients, N_{ASA}, can at the most be equal to the number of the alternatives less one, i.e. $N_{ASA} \leq (J-1)$. In this case the LR statistic becomes:

$$LR(\beta_{\mathrm{ASA}}) = -2\ [ln\ L(\beta^{ML}{}_{ASA}) - ln\ L\ (\beta^{ML})] \tag{8.3.16}$$

Fig. 8.3.4 shows the statistics $LR(\mathbf{0})$ and $LR(\beta_{ASA})$ with their respective degrees of freedom. These statistics far exceed the 95th percentile of the corresponding chi-square variables with N_β and $N_\beta - N_{ASA}$ degrees of freedom and therefore the assumptions that the "true" model has null coefficients or has only modal constants can be rejected with a very low error probability.

Statistics and tests on goodness of fit. The model's capability to reproduce the choices made by a sample of users[7] can be measured by using the rho-square statistic:

$$\rho^2 = 1 - \frac{lnL(\beta^{ML})}{lnL(\mathbf{0})} \tag{8.3.17}$$

This statistic is a normalized measure in the interval [0,1]. It is equal to zero if $L(\beta^{ML})$ is equal to $L(\mathbf{0})$, i.e. the model has no explanatory capability; it is equal to one if the model gives a probability equal to one of observing the choices actually made by each user in the sample, i.e. the model has perfect capability to reproduce observed choices.

Alternatively, it is possible to use an adjusted value of rho-square statistic, sometimes named rho-square bar, which substitutes the log-likelihood function lnL (β^{ML}) with its unbiased estimate lnL (β^{ML})- N_β, where N_β is the number of parameters estimated in the model:

$$\bar{\rho}^2 = 1 - \frac{lnL\left(\beta^{ML}\right) - N_\beta}{lnL(\mathbf{0})} \tag{8.3.18}$$

Expression (8.3.18) attempts to eliminate the effect of the number of parameters included in the model's specification to allow the comparison of models with different numbers of parameters.

The adjusted rho-square statistic can be used, in fact, to compare two models (model 1 and model 2) which are specified in different ways, i.e. such that the vectors β_1 and β_2 can not be obtained as a special case of the other[8]. In this case, under the null hypothesis that model 1 is "true", the probability that the statistic $\bar{\rho}_2^2$

of model 2 is for sampling reasons larger by some z than that of model 1, is inferior to the value of the probability distribution function of a Standard Normal variable, $N(0;1)$, computed for the value

$$\bar{z} = - [-2\, z\, ln\, L(\mathbf{0}) + (N_1 - N_2)]^{1/2} \tag{8.3.19}$$

or

$$Pr(\bar{\rho}_2^2 - \bar{\rho}_1^2 > z) \leq \phi(\bar{z}); z > 0 \tag{8.3.20}$$

where $\phi(\bar{z})$ is the value of the p.d.f. of $N(0.1)$ and N_1 and N_2 are the number of parameters in model 1 and 2 respectively.

In addition to the statistics ρ^2 and $\bar{\rho}^2$, other informal statistics are used to assess qualitatively the goodness of fit of a model. One of these statistics (% *right*) relates to the percentage of observations in the sample for which the alternative actually chosen is that of maximum probability as predicted by the model. Other synthetic statistics are the choice percentage observed and predicted by the model for each alternative. The former is given by the ratio between the number of users choosing each alternative and the total number of users to whom it is available. The latter is obtained as the average of choice probabilities given by the model for the users to whom the alternative is available.

Tests on the functional form. The statistical tests described above are relative to different hypotheses on the coefficients β^{ML} obtained from the calibration of a model, assuming its specification as given. This section describes some statistical tests that compare different hypotheses on the functional form of the model.

Two generic alternative specifications can be compared by using the $\bar{\rho}^2$ test in equation (8.3.20). Alternatively, specific tests related to particular functional forms can be used. For example, in Chapter 3 it was shown that the Multinomial Logit model is a special case of a single-level Hierarchical Logit if δ=1 (expression (3.2.24)), and of the multi-level Hierarchical Logit, if δ_r=1 for each intermediate node r of the choice tree (section 3.3.3). The hypothesis that the "true" model is a Multinomial Logit can be tested by calibrating Hierarchical Logit models and testing the null hypothesis that the estimates δ^{ML} are equal to one. These tests can be conducted using the statistics described previously for testing hypotheses on single or multiple parameters. For the Multinomial Logit model, the Independence of Irrelevant Alternatives (IIA) property discussed in section (3.3.1) can be tested directly.

Under the IIA hypothesis, the choice model for any subset I' of alternatives, (partial choice set) contained in I, (universal choice set) $I' \subseteq I$, is still a Multinomial Logit model:

$$p^i[j/I'] = exp(\bar{\beta}^T X^i_j) / \sum_{h \in I'} exp\, \bar{\beta}^T X^i_h \qquad (8.3.21)$$

where $\bar{\beta}$ indicates the sub-vector of coefficients included in the systematic utilities of the alternatives contained in I' (e.g. $\bar{\beta}$ will not contain the coefficients of the ASA of alternatives not belonging to I'). The number of these coefficients will be $N_{\bar{\beta}} \leq N_{\beta}$. The Maximum Likelihood estimator $\bar{\beta}^{ML}_{I'}$ for the model (8.3.21) can be obtained on the sub-sample of observations choosing the alternatives in I'. If the IIA hypothesis is true, the vector $\bar{\beta}^{ML}_{I}$ of the $N_{\bar{\beta}}$ coefficients obtained by calibrating the model for all the alternatives over the whole sample and the vector $\bar{\beta}^{ML}_{I'}$ described previously must be statistically equivalent. This hypothesis can be tested by using the statistic:

$$\left(\bar{\beta}^{ML}_{I} - \bar{\beta}^{ML}_{I'}\right)^{T} \left(\Sigma_{\bar{\beta}_I} - \Sigma_{\beta_{I'}}\right)^{-1} \left(\bar{\beta}^{ML}_{I} - \bar{\beta}^{ML}_{I'}\right) \qquad (8.3.22)$$

which under the null hypothesis is distributed according to a chi-square variable with $N_{\bar{\beta}}$ degrees of freedom. The matrices Σ_{β_I} and $\sum_{\beta_{I'}}$ are the variance-covariance matrices of the estimates $\bar{\beta}^{ML}_{I}$ and $\bar{\beta}^{ML}_{I'}$ of the $N_{\bar{\beta}}$ common components. To test the IIA hypothesis, the test should be carried out on different subsets I' of the universal choice set I.

8.4. Disaggregate estimation of demand models with Stated Preferences surveys*

The information on travel behavior needed to specify and calibrate demand models can also be obtained using *Stated Preference* (*SP*) surveys. This term refers to a set of techniques using the statements made by interviewees about their preferences in hypothetical scenarios. SP techniques are based on the possibility of "controlling the experiment" by designing the choice context rather than recording choices in a given, generally uncontrolled, choice context as in the case with *Revealed Preference* (RP) surveys described in the previews section. SP surveys have several advantages over RP surveys, which can be summarized as follows:

- they allow the introduction of choice alternatives not available at the time of the survey (e.g. new modes or services in a mode choice context);
- they can control the variation of relevant attributes outside the present range to obtain better estimates of the relative coefficients. For example, the monetary cost in urban areas is usually contained within a limited range;
- they can introduce new attributes not present in the real choice context (e.g. information to the passengers, vehicle air-conditioning, other on-board services);

- they can collect more information, i.e. larger samples, per unit cost since each interviewee is usually asked about several choice contexts.

These advantages are obtained at a price of introducing some distortion in the results and in the models calibrated. Distortions stem from the possible differences between stated and real choice behavior; if the user experienced a real situation, his/her behavior might be different from that stated during the SP survey. Differences in behavior may be due to several factors. For example, the context suggested might be or appear to be unrealistic, some attributes of the suggested alternative relevant for the decision maker might be missing, there may be fatigue and justification bias effects. The analysis of the possible causes of distortion and of the remedies is outside the scope of this book. However, it should be noted that some of these problems are structural, or ingrained in SP surveys technique, while others can be solved by careful design and execution of the surveys bringing them as close as possible to real choice contexts.

From the above, it is clear that SP surveys, in spite of their considerable application potential, should be seen as complementary, rather than alternative, to RP techniques. The advantages and disadvantages of the two techniques compensate each other and, as will be seen, the techniques can be used jointly to build demand models.

In practice, in the field of SP techniques there are several different approaches appropriate for different aims. In the following, reference will be made to the SP techniques most widely used for the specification and calibration of travel demand models. In particular, section 8.4.1 will introduce some definitions and the main types of surveys, section 8.4.2 will describe some aspects of SP survey design, while section 8.4.3 will deal with model calibration methods using the combined results of RP and SP surveys.

8.4.1. Definitions and types of survey

A Stated Preferences experiment is fully identified by a number of elements: the composition of the choice contexts proposed to the decision maker, the selection of the choice contexts proposed, the type of preference elicited from the decision maker and the way in which the interview is conducted.

During the interview, the decision-maker is usually presented with different *scenarios* or *choice contexts*. A scenario is defined by the set of *alternative options*[9]; each option is accompanied with some *attributes* or *factors* defining its characteristics. Fig. 8.4.1 shows two choice contexts (scenario A and scenario B), each consists of two alternative modes and their attributes.

In the choice contexts proposed, the attributes vary between a prefixed number of values, or *levels*. These levels can be defined in absolute terms, e.g. travel times and costs, or obtained as percentage variations with respect to the values of the attributes for a real context experienced or known to the decision maker (e.g. times and costs relative to certain origin-destination pairs).

The decision-maker can be asked about different types of *preference*:

- *choice*, i.e. an indication of which option he/she would choose in that context;
- *ranking*, i.e. a ranking of the available options according to his/her preferences;
- *rating*, i.e. the assignment of a vote of preference on a predefined scale for each alternative option.

Note that the three types of preference provide a gradually increasing quantity of information but require an increasing involvement of the decision-maker. Furthermore, "choice" and "ranking" coincide when the choice context consists of only two alternative options.

The number of possible scenarios depends on the number of combinations of the design elements introduced, namely the number of options, the number of attributes and the number of levels for each attribute. Since the total number of scenarios might be too large and not all the combinations are equally "useful", one of the elements in the design of an SP experiment will be the *selection* of the scenarios to be proposed to the decision maker/s.

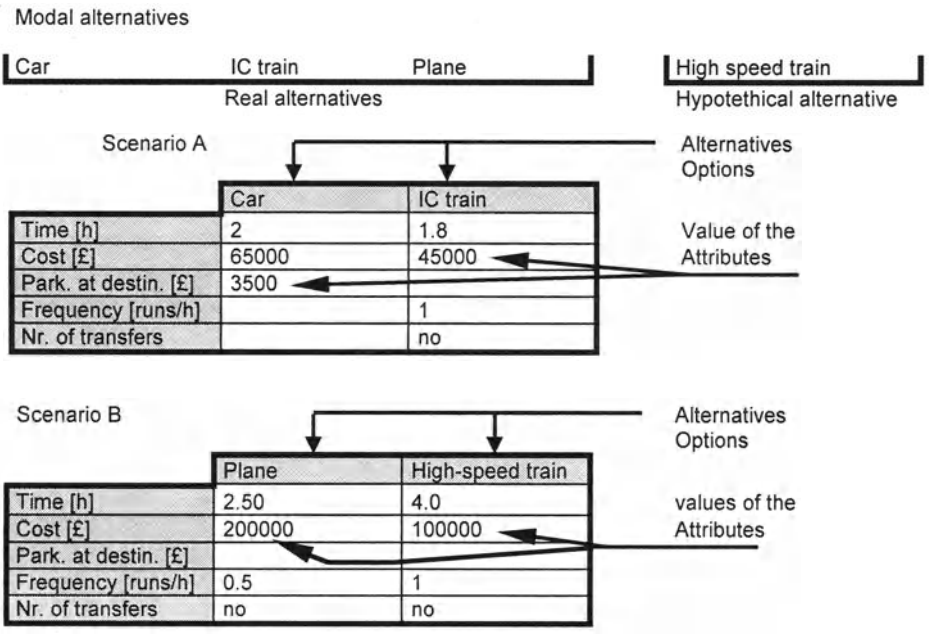


Fig. 8.4.1 Hypothetical scenarios for an SP survey

Finally, the interviews can be conducted using different procedures. In traditional methods, the decision-maker is asked to fill in pre-printed paper forms. In

more sophisticated computer-aided techniques, the scenarios are generated in real-time taking into account previous answers.

8.4.2. Survey design

Designing an SP survey requires the definition of all the elements described. It must be remembered that, in spite of the operational guidelines and the theoretical analyses, SP survey design, even more than with traditional surveys, is a synthesis based on the analyst's experience and sensitivity. The main operational suggestions resulting from many years of research and experimentation are summarized below.

- *Scenario realism*: results of SP surveys are significantly better if choice scenarios are in the direct experience of the decision-maker. For example, in a survey for the calibration of a modal choice model, an RP interview can be carried out on an actual journey of a certain type, then SP scenarios obtained from that journey by varying the attributes or by introducing a new mode can be proposed. In this way the distortion effects described above can be reduced considerably. It is obvious that this type of survey requires more preparation; portable computers can generate in real time level-of-service attributes for the different modes.
- *"Choice" rather than "ranking" and "rating"*. It seems that greater simplicity and less ambiguity of preference statements compensate for the smaller amount of information produced by this type of experiment. In addition it is possible to use results and estimation techniques analogous to those obtained for RP surveys.
- *Scenario simplification.* It seems that proposing a limited number of alternative options defined by a reduced number of attributes gives rise to better results.
- *Limitation of the number of scenarios* proposed to each decision-maker in order to avoid fatigue effects deteriorating the quality of results. Many experiences suggest that each decision-maker should be confronted with no more than nine or ten scenarios.

The latter aspect is strictly connected to the most theoretical phase of survey design, namely scenario selection. In most cases the number of scenarios theoretically possible is very large; in fact, subdividing the n factors in k groups (i = 1,2,...,k) of n_i elements taking on m_i levels, the total number N of possible scenarios will be:

$$N = \prod_{i=1}^{k} m_i^{n_i}$$

The number of factors must be computed taking into account that any single attribute present in p alternatives counts for p different factors.

A *Full Factorial Design* considers all possible scenarios. There are many techniques[10] for reducing the number of scenarios in a Full Factorial Design, generating a subset of scenarios with same desirable properties. Some results for the case of two levels per factor are given below; the case of several levels can be reduced by decomposing a multi-level factor into many two-level factors and introducing some compatibility constraints on the combinations of levels that the new factors can assume.

Fig. 8.4.2 lists all the possible scenarios, indicating with + and – the two levels of each factor for an experiment with three factors and two levels (N=8); factors are time and cost for the car (*TC* and *CC*) and time for the bus (*TB*).

SCENARIO NR.	AVERAGE	FACTORS			INTERACTIONS				RESULT OF CHOICE
		TC	*CC*	*TB*	*TC,CC*	*TC,TB*	*CC,TB*	*TC,CC,TB*	
1	+	-	-	-	+	+	+	-	U_1
2	+	+	-	-	-	-	+	+	U_2
3	+	-	+	-	-	+	-	+	U_3
4	+	+	+	-	+	-	-	-	U_4
5	+	-	-	+	+	-	-	+	U_5
6	+	+	-	+	-	+	-	-	U_6
7	+	-	+	+	-	-	+	-	U_7
8	+	+	+	+	+	+	+	+	U_8
Divisor	8	4	4	4	4	4	4	4	

Fig. 8.4.2 Example of Full Factorial Design with levels and main interaction effects.

It will also be assumed that the experiment associated with the i-th scenario (i-th row of the matrix in Fig. 8.4.2) yields an observation of the variable U_i not known a priori. In the example this variable could be the difference of the perceived utility between the two alternatives (rating), or a binary indicator of the alternative preferred by the decision maker (ranking and choice). Furthermore, l_{ij} indicates the level of the jth factor in the ith scenario, or the generic element of the matrix in Fig.8.4.2, which under the assumptions made assumes the values +1 and –1 in correspondence with the "high" and "low" level of the factor. The complete experiment is defined a "comparison" for the factor j if it results:

$$\sum_{i=1,\ldots,N} l_{ij} = 0 \tag{8.4.1}$$

i.e., if the number of high levels (+) is equal to the number of low levels (-) in the N scenarios making up the experiment. Two comparisons relative to the factors j and h are said to be "orthogonal" if:

$$\sum_{i=1,\ldots,N} l_{ij}\ l_{ih} = 0 \tag{8.4.2}$$

i. e., if the numbers of scenarios in which the levels of the two factors are concordant (++,--), is equal to the number in which they are discordant (+-, -+).

The variation (total variance) of the variables U_i can be explained in terms of the "main effects" and of "interaction effects" of the factors considered in the experiment.

The main effect of the factor j, $P_{(j)}$, is defined as the difference between the two averages $\overline{U}_+$ and $\overline{U}_-$ of the variable U calculated respectively in correspondence of the values (+) and (-) of the factor. If the vector $\boldsymbol{l}_j$ is a comparison, it therefore results:

$$P_{(j)} = \frac{2}{N} \sum_{i=1,\ldots,N} l_{ij} U_i \tag{8.4.3}$$

For the example in Fig. 8.4.2, the main effect of the factor *TC* is therefore:

$$P_{(TC)} = \frac{1}{4}(U_2 + U_4 + U_6 + U_8) - \frac{1}{4}(U_1 + U_3 + U_5 + U_7)$$

The interaction effect between the factors j and h, $I_{(j,h)}$, is defined as the difference between the averages of the variable U obtained for the concordant values, (+) (+) or (-) (-), and the discordant values, (+) (-) or (+) (-), of the two factors. If the two vectors $\boldsymbol{l}_j$ and $\boldsymbol{l}_h$ are comparisons, it results:

$$I_{(j,h)} = \frac{2}{N} \sum_{i=1,\ldots,N} l_{ij} l_{ih} U_i \tag{8.4.4}$$

For the example of Fig. 8.4.2, the interaction effect *TC, CC* is therefore:

$$I_{(T_C, C_C)} = \frac{1}{4}(U_1 + U_4 + U_5 + U_8) - \frac{1}{4}(U_2 + U_3 + U_6 + U_7)$$

Furthermore, from (8.4.4) it is deduced that, analogously to the levels of a factor, the level of interaction between two factors (j,h) for the i-th scenario, $l_{i(j,h)}$, can be defined as $l_{ij} \cdot l_{ih}$; the interaction effect between the two factors can therefore be expressed as:

$$I_{(j,h)} = \frac{2}{N} \sum l_{i(j,h)} U_i \tag{8.4.5}$$

Analogously, the interaction effect of three factors (j, h, k) can be defined as:

$$I_{(j,h,k)} = \frac{2}{N} \sum_{i=1,\ldots,N} l_{ij} l_{ih} l_{ik} U_i \qquad (8.4.6)$$

and the interaction level of three factors can be expressed as $l_{i(j,h,k)}$.

Fig. 8.4.2 shows the two-factor interaction levels and the unique three-factors interaction level as well as the average column, I, made up of the variables m_i all equal to (+1). These variables allow the expression of the average of U as:

$$\bar{U} = \frac{1}{N} \sum_{i=1,\ldots,N} m_i U_i$$

Under the assumption of orthogonal comparisons, the N values assumed by the variable U can be entirely explained as a linear combination with coefficients a_i of the average, of the main effects and of the interaction effects between the different factors. In the case of the example in Fig. 8.4.2, we have:

$$U_i = \alpha_1 m_i + \alpha_2 l_{i(TC)} + \alpha_3 l_{2(CC)} + \alpha_4 l_{i(TB)} + \alpha_5 l_{i(TC\,CC)} + \\ + \alpha_6 l_{i(TC\,TB)} + \alpha_7 l_{i(CC\,TB)} + \alpha_8 l_{i(TC\,CC\,TB)}$$

Many experiments, however, lead to the conclusion that most of the overall variance of the variable U is explained by the main effects (approximately 80%), while the two-factor interaction effects explain a limited fraction of the global variance (3-6%). Furthermore, the variance explained by the interactions of more than two factors is usually negligible. In other words, if the variable U were expressed as a linear combination of the average and of the main effects, the variance explained by the model would be around 80% of the total variance observed for the variable U and so on. Starting from these results, it is possible to introduce some partialization techniques of the experiment. Examples are techniques to reduce the number ($N' < N$) of scenarios presented to each decision-maker, while retaining the orthogonality of the comparisons and the possibility of evaluating at least the main effects of the factors considered.

The first technique, known as "*block decomposition*" of the Full Factorial Design is based on the principle of subdividing the set of alternative scenarios in groups (blocks) to present to different decision-makers. In order to obtain blocks satisfying the properties (8.4.1) (comparisons) and (8.4.2) (orthogonality between comparisons) one or more "block variables" are selected and the scenarios corresponding to the same value of the block variable, or concordant (discordant) values of many block variables, are grouped together. The block variables normally used are high-level interactions since the effects on the variance of the block variables and their interactions can be estimated only approximately on the basis of the observation of all the interviewees. Fig. 8.4.3 shows two subdivisions into blocks of the full design in Fig. 8.4.2. In the first case, the eight scenarios have been divided into two blocks of 4 (8/2) scenarios, using the interaction level of the three

factors (*TC*, *CC TB*) as the block variable. In the second case, 4 blocks (8/2x2) of two scenarios each have been obtained using as block variables the interactions level of the two factors (*TC*, *CC*) and of the two factors (*TC*, *TB*).

Another partialization technique of the Full Factorial Design, known as *Fractional Factorial Design*, eliminates completely some scenarios while retaining orthogonal comparisons which allow the estimation of the main effects. If the resulting number of scenarios is still too high to be presented to a single decision-maker, they can be further broken down into blocks by using the method described previously. A fractional factorial design can be obtained from the full design through a "defining relationship". The simplest case is that in which the level of a given factor is obtained from those of all the others resulting from a full design which excludes the factor to be obtained. The level of the "derived" factor is supposed equal to the level of a higher-level interaction effect. For example, in the case of Fig. 8.4.4 it is assumed that the level of the factor *TB* is equal to that of the interaction effect (*TC*, *CC*), where the levels of *TC* and *CC* are those defined in a two-factor full design ($N=2^2$); in this case the following "defining relationship" has been adopted:

$$TB = (TC,\ CC) \text{ i.e. } l_{iTB} = l_{iTC} \cdot l_{iCC} \tag{8.4.7}$$

SCE. NR.	FACTORS			BLOCK VAR.	ALTERNATIVES ORGANIZATED IN BLOCK					SC.. NR.
	TC	*CC*	*TB*	*TC,CC,TB*	Block		*TC*	*CC*	*TB*	
1	-	-	-	-	I	Block I	-	-	-	1
2	+	-	-	+	II		+	+	-	4
3	-	+	-	+	II		+	-	+	6
4	+	+	-	-	I		-	+	+	7
5	-	-	+	+	II	Block II	+	-	-	2
6	+	-	+	-	I		-	+	-	3
7	-	+	+	-	I		-	-	+	5
8	+	+	+	+	II		+	+	+	8

SC. NR	FACTORS			BLOCK VAR.		ALTERNATIVES ORGANIZATED IN BLOCK					SC. NR
	TC	*CC*	*TB*	*TC, CC*	*TC, TB*	block		*TC*	*CC*	*TB*	
1	-	-	-	+	+	IV	Block I	+	-	-	2
2	+	-	-	-	-	I		-	+	+	7
3	-	+	-	-	+	II	Block II	-	+	-	3
4	+	+	-	+	-	III		+	-	+	6
5	-	-	+	+	-	III	Block III	+	+	-	4
6	+	-	+	-	+	II		-	-	+	5
7	-	+	+	-	-	I	Block IV	-	-	-	1
8	+	+	+	+	+	IV		+	+	+	8

Fig. 8.4.3 Construction of two and four blocks from the Full Factorial Design in Fig. 8.4.2

The design in Fig.8.4.4 is thus obtained as follows:

- development of the full factorial design for the two factors *TC* and *CC*;
- calculation of the interaction effect level (*TC*, *CC*);
- definition of the level of factor *TB* using the equation (8.4.7).

SCENARIO	FACTORS		INTERACTION	FACTOR
NR.	*TC*	*CC*	*TC*, *CC*	*TB*
1	-	-	+	+
2	+	-	-	-
3	-	+	-	-
4	+	+	+	+

Fig. 8.4.4 Example of Fractional Factorial Design for the Full Factorial Design in Fig. 8.4.2.

With a fractional factorial design the possibility of estimating some interaction effects is lost, as these effects get "confused" with the "retained" ones. Confused effects can be identified by manipulating the "defining relationship" of the fractional factorial design. Thus, recalling that the product of the levels of the same factor is equal to the average factor *I*, for the relationship (8.4.7) it results:

$$\begin{aligned} TB \times TB &= TB \times TC \times CC = I \\ TC \times TB &= TC \times TC \times CC = CC \\ CC \times TB &= CC \times CC \times TC = TC \end{aligned} \qquad (8.4.8)$$

i.e., the three-factor interaction effect (*TC*, *CC*, *TB*) gets confused with the average and the two-factor interaction effects (*TC*, *TB*) and (*CC*, *TB*) get confused with the primary effects of the factors *CC* and *TC* respectively. Obviously the two-factor interaction effect (*TC*, *CC*) is confused with the primary effect *TB* by construction.

The "length" of the defining relationship, i.e. the number of factors in it, is known as the resolution of a fractional factorial design. The resolution of equation (8.4.7) is equal to 3. The number of scenarios in a fractional factorial design depends on the number of defining relationships; for each defining equation, under the assumption of two levels for each factor, the number of scenarios halves. Obviously the choice of defining relationships must be based on the analyst's expectations concerning the relevant effects which are not to be confused in order to explain the observed behaviors.

To give a more detailed example, suppose there are seven factors generically indicated by *A*, *B*, *C*, *D*, *E*, *F*, *G* with two levels each. The full factorial design has $2^7 = 128$ scenarios, a fractional factorial design with $2^{7-1} = 64$ scenarios can be obtained with a single defining relationship. For example:

$$G = (A\ B\ C\ D\ E\ F)$$

A design with $2^{7-4} = 8$ scenarios can be obtained with 4 defining relationships, for example:

$$D = (A,B);\ E = (A,C);\ F = (B,C);\ G = (A,B,C)$$

and so on.

Note the difference between the two methods described for partializing the full factorial design. With the block variables method, the whole full factorial design is used, even if it is presented to several decision-makers; with the fractional design, on the other hand, some scenarios are completely eliminated. In the former case many scenarios are generated but, given the number of decision-makers in the sample, less information (preference statements) is obtained for each scenario; with the fractional factorial design, the opposite occurs.

It should be pointed out that SP surveys are often aimed at the calibration of random utility models whose systematic utility function usually include the values of individual attributes or their functional transformation. This is equivalent to the assumption that the interactions between the attributes (or factors) can be disregarded in explaining the choice behaviors of decision-makers. Thus, the SP survey design should allow at least the evaluation of all the main effects of the factors considered.

8.4.3. Model calibration

The results of a SP survey can be used to calibrate demand models relative to the choice dimensions proposed to the decision-makers. The estimation methods used in practice are analogous to those described for Revealed Preferences in section 8.3.2. In fact, each scenario i presented to a decision-maker can be seen as an element of a sample of observations of choice behaviors. The final size of the SP sample thus is equal to:

$$n_{SP} = \sum_{z=1,\ldots,N_{SP}} n_z$$

where n_z is the number of scenarios presented to the zth decision maker and N_{SP} is the number of decision makers contemplated in the SP survey.

Furthermore, the attributes proposed for the different alternatives and the chosen alternative, $j(i)$, can be associated to each scenario i. The chosen alternative is the alternative explicitly chosen by the decision maker in "choice" surveys, or the one with greatest attractiveness in "ranking" or "rating" surveys. Under the approximate assumption that the n_{SP} observations are statistically independent[11], it is possible to formulate likelihood and log-likelihood functions for the SP sample formally coincident with expressions (8.3.2) and (8.3.3) and all the results described there can be extended to the estimation of SP-based models.

As stated in the introduction, SP surveys should be considered as complementary to traditional RP surveys and the combined use of the two can balance reciprocal merits and shortcomings. From the point of view of demand modeling it is therefore useful to carry out *joint calibrations* using RP and SP surveys on the same sample or on different samples of users. Random utility models explaining RP and SP choices should be specified separately since their attributes, random residuals variances and, in principle, even functional forms might be different. Possible specifications of the perceived utilities in both models are formalized below:

RP MODEL:

$$U_{ji}^{RP} = \boldsymbol{\beta}^{T} \boldsymbol{X}_{ji}^{RP} + \eta^{T} \boldsymbol{W}_{ji}^{RP} + \varepsilon_{ji}^{RP} = V_{ji}^{RP} + \varepsilon_{ji}^{RP} \qquad (8.4.9)$$

$$i = 1,\ldots, n_{RP}$$

where:

U_{ji}^{RP} is the perceived utility associated with alternative j by decision maker i in the RP context;

$\boldsymbol{X}_{ji}^{RP}$ is the vector of the common attributes relative to the alternative j for decision maker i; these attributes appear in the specification of the SP model with the same coefficients;

$\boldsymbol{W}_{ji}^{RP}$ is the vector of the RP specific attributes relative to alternative j for decision maker i;

ε_{ji}^{RP} is the random residual of alternative j for decision maker i ;

V_{ji}^{RP} Is the systematic utility of the RP model associated with alternative j, for decision maker i;

$\boldsymbol{\beta}$ and η are the vectors of the unknown parameters to be estimated.

SP MODEL:

$$U_{ji}^{SP} = \boldsymbol{\beta}^{T} \boldsymbol{X}_{ji}^{SP} + \gamma^{T} \boldsymbol{Z}_{ji}^{SP} + \varepsilon_{ji}^{SP} = V_{ji}^{SP} + \varepsilon_{ji}^{SP} \qquad (8.4.10)$$

$$i = 1,\ldots, n_{SP}$$

where:

U_{ji}^{SP} is the perceived utility associated with alternative j in the hypothetical scenario i;

$\boldsymbol{X}_{ji}^{SP}$ is the vector of the common attributes relative to alternative j for the scenario i;

$\boldsymbol{Z}_{ji}^{SP}$ is the vector of the SP specific attributes relative to alternative j for the scenario i;

ε_{ji}^{SP} is the random residual of alternative j for the scenario i ;

V_{ji}^{SP} is the systematic utility of the SP model associated with alternative j for scenario i;

β and γ are the vectors of the unknown parameters to be estimated.

The definition of the choice probabilities $p_{RP}^{i}[j]$e $p_{SP}^{i}[j]$ obviously depends on the assumptions on the distribution of the random vectors $\boldsymbol{\varepsilon}^{RP}$ and $\boldsymbol{\varepsilon}^{SP}$. Assuming ε_{ji}^{SP} and ε_{ji}^{RP} as i.i.d. Gumbel variables of parameters θ_{SP} and θ_{RP} respectively, the probability of choosing alternative j in the observation (decision maker or scenario) i assumes the form of a Multinomial Logit model for both the RP and the SP models:

$$p_{RP}^{i}[j]=\frac{exp(V^{RP}{}_{ji}/\theta_{RP})}{\sum_{h} exp(V^{RP}{}_{hi}/\theta_{RP})}; \qquad p_{SP}^{i}[j]=\frac{exp(V^{SP}{}_{ji}/\theta_{SP})}{\sum_{h} exp(V^{SP}{}_{hi}/\theta_{SP})} \tag{8.4.11}$$

Specific attributes of the SP model may include qualitative attributes, such as on-board comfort, services contemplated in the SP survey or ASA for alternatives non-available in the RP context (e.g. new transport modes or services).

State dependence or state inertia is one of the SP specific attributes often included in the specification (8.4.10). This attribute represents the conditioning of the generic SP decision-maker with respect to the alternative actually chosen (RP context). Inertia is usually modeled as a dummy variable equal to one if the user i chooses alternative j in the RP context, zero otherwise. Its coefficient is usually statistically significant and positive to indicate, given the values of other attributes, a larger perceived utility and choice probability for the alternative chosen in the real context. Obviously the state dependence attribute can be used only if the RP and SP surveys relate to the same sample of decision-makers.

Specific attributes of the RP model may be relative to variables not included among the SP factors.

A scale factor taking into account the possibility that the variances of the vectors $\boldsymbol{\varepsilon}^{RP}$ and $\boldsymbol{\varepsilon}^{SP}$ might be different is usually introduced for joint calibration. In fact, as stated in section 8.3., for models belonging to the Logit family it is not possible to estimate the parameter θ separately from the coefficients β_k and the estimates $\hat{\beta}_k^{ML}$ are in reality ratios $\beta_k'=\beta_k/\theta$. To take into account the possible difference of the

variances of the residuals ε^{RP} and ε^{SP}, a scale factor μ, equal to the ratio between the parameters θ of the two random vectors, is introduced:

$$\mu^2 = \frac{Var[\varepsilon_{RP}]}{Var[\varepsilon_{SP}]} = \frac{\theta_{RP}^2}{\theta_{SP}^2} \quad \text{i.e.} \quad \mu = \frac{\theta_{RP}}{\theta_{SP}} \tag{8.4.12}$$

The log-likelihood function for the RP and SP samples can therefore be expressed including the parameter θ_{RP} in all the other coefficients:

$$lnL^{RP}(\beta', \eta') = \sum_{i=1,\ldots,n_{RP}} ln\, p_{RP}[j(i)]\left(X_i^{RP}, W_i^{RP}, \beta', \eta'\right) \tag{8.4.13}$$

$$lnL^{SP}(\beta', \gamma', \mu') = \sum_{i=1,\ldots,n_{SP}} ln\, p_{SP}[j(i)]\left(X_i^{SP}, Z_i^{SP}, \beta', \gamma', \mu'\right) \tag{8.4.14}$$

where the probabilities $p[j(i)]$ are obtained by using the following systematic utilities:

$$V_{ij}^{RP} = \beta'^T X_{ij}^{RP} + \eta'^T W_{ij}^{RP} \qquad \beta' = \beta / \theta_{RP} \quad \eta' = \eta / \theta_{RP}$$
$$V_{ij}^{SP} = \mu\beta'^T X_{ij}^{SP} + \gamma'^T Z_{ij}^{SP} \qquad \gamma' = \gamma / \theta_{SP} \quad \mu\beta' = \beta / \theta_{SP}$$

The combined estimate of the parameters (β', η', γ', μ) can therefore be obtained by maximizing the log-likelihood function of the joint sample which is the sum of expressions (8.4.13) and (8.4.14) on the assumption that the RP and SP samples are independent:

$$\begin{aligned}(\beta', \eta', \gamma', \mu') &= arg\, max\left[ln\, L^{RP+SP}(\beta', \eta', \gamma', \mu)\right] = \\ &= arg\, max\left[ln\, L^{RP}(\beta', \eta') + ln\, L^{SP}(\beta', \gamma', \mu)\right]\end{aligned} \tag{8.4.15}$$

Note also that under the hypothesis that the two choice models $p_{RP}[j(i)]$ and $p_{SP}[j(i)]$ are Multinomial Logit such as (8.4.11), the global log-likelihood function (8.4.15) is concave in the parameters β', η' and γ', but not in the scale factor μ as it is the case with the Hierarchical Logit model. This implies that for the maximization of (8.4.15), it is not possible to use the same gradient algorithms described in section 8.3.2 for the Multinomial Logit model. A possible solution is to use the gradient algorithms to maximize the function $lnL(\beta', \delta', \gamma', \mu^k)$ for a predefined value μ^k of the scale factor, associated with a mono-dimensional optimization algorithm that explores different values of μ (see Appendix A).

Experimental evidence indicates that the combined use of RP and SP data for estimating the parameters usually results in an improvement in statistical precision

and in more reasonable parameter values. Furthermore, it is not possible to define a priori whether the scale factor μ must be greater or less than one. In fact, there are reasons both for a larger variance of RP random residuals (less precise attributes used for calibration, omitted attributes, etc.) and for the opposite (less realism of the choice context, fatigue effect, etc.). As an example, Fig. 8.4.5 reports the results of the calibrations of a Multinomial Logit mode choice model using RP and SP data separately and jointly.

PARAMETERS (ATTRIBUTES)	RP MODEL	SP MODEL	RP/SP MODEL
β_1 (travel time)	-3.277 (-2.2)	-2.585 (-3.9)	-2.82 (-3.9)
β_2 (cost)	-2.863 (-3.5)	-1.336 (-5.9)	-1.371 (-2.4)
V.O.T	1144 £/h	1934 £/h	2056 £/h
β_3 (access time)	-6.606 (-1.2)	-3.176 (-3.5)	-4.776 (-4.8)
β_4 (waiting time)	-10.40 (-2.3)	-19.62 (-4.1)	-20.86 (-4.0)
β_5 (Nr of motorbikes)	5.391 (3.9)	2.831 (3.6)	2.848 (5.5)
β_6 (Nr. of cars)	3.175 (2.5)	1.933 (4.4)	1.528 (3.9)
β_7 (chain)	-1.399 (-1.2)	-1.730 (-2.3)	-.4545 (-1.7)
η_1 (ASA Car RP)	-1.370 (-1.2)		-4.271 (-4.9)
η_2(ASA Motorbike RP)	-4.492 (-3.3)		-6.076 (-6.8)
γ_1 (ASA car SP)		-9748 (-1.8)	-1.923 (-3.1)
γ_2(ASA Motorbikes SP)		-1.499 (-2.3)	-2.480 (-3.4)
γ_3 (Inertia)			2.603 (4.4)
Scale factor μ			0.786 (4.0)
	STATISTICS		
ln L(0)	-105.4668	-408.6838	-514.1506
Ln L(β)	-55.4268	-210.4182	-282.2376
LR	100.08	396.53	463.826
RHO2	.4745	.4851	.6612

RP model

$$V^{RP}_{car} = \beta_1 Tb + \beta_2 Mc + \beta_6 Nc \beta_7 CH + \eta_1 CAR^{RP}$$
$$V^{RP}_{Motorbike} = \beta_1 Tb + \beta_2 Mc + \beta_5 Nm + \eta_2 MOTORBIKE^{RP}$$
$$V^{RP}_{Bus} = \beta_1 Tb + \beta_2 Mc + \beta_3 Ta + \beta_4 Tw$$

SP model

$$V^{SP}_{car} = \beta_1 Tb + \beta_2 Mc + \beta_6 Nc\ \beta_7 CH + \gamma_1 CAR^{SP} + \gamma_3 IN$$
$$V^{SP}_{Motorbike} = \beta_1 Tb + \beta_2 Mc + \beta_5 Nm + \gamma_2 MOTORBIKE^{SP} + \gamma_3 IN$$
$$V^{SP}_{Bus} = \beta_1 Tb + \beta_2 Mc + \beta_3 Ta + \beta_4 Tw + \gamma_3 IN$$

Attributes

Tb	=	Travel time on board [*h*]
Mc	=	Monetary Cost [£·10^3]
Nm,Nc	=	Nr. of motorbikes and cars in household
Ta	=	Access time [*h*]
Tw	=	Waiting time[*h*]
CH	=	Dummy variable (0/1), 1 if the trip belongs to a chain (sequence of more than 2 trips)
Car, Motorbike	=	Alternative Specific Attributes (ASA)
IN	=	Inertia Variable (0/1), 1 if mode was chosen in the RP survey

Fig. 8.4.5 Separate and joint RP/SP calibrations of a Multinomial Logit mode choice model

The joint calibration on RP and SP data of more complex random utility models is further complicated by the parameters defining the joint density function of the vectors ε^{RP} and ε^{SP} which are generally more than one. If the two models $p_{RP}[j]$ and $p_{SP}[j]$ were two Hierarchical Logit models with the same tree structure (the same vector of parameters δ), it would still be possible to introduce a scale factor μ relative to the variances of the two random vectors. For different correlation structures or other functional forms it would be more complicated to synthesize the different structure of the variance-covariance matrices of ε^{RP} and ε^{SP} with few parameters.

8.5. Estimation of O-D demand flows using traffic counts

This section covers the methods aimed at improving the estimates of present origin-destination demand flows by combining direct and/or indirect (model) estimators with other aggregate information related to O-D demand flows. In the following the aggregate information will be identified with traffic counts, i.e. counts of user flows, on some elements (links) of the transportation supply system (network)[12]. The problem of estimating O-D flows by combining traffic counts with all the other available information in the literature is sometimes referred to as origin-destination Count Based Estimation (ODCBE) problem.

From a certain point of view, the problem of estimating O-D flows by using traffic counts can be considered as the inverse assignment problem. The latter in Chapter 5 was stated as that of calculating link flows starting from O-D flows, network and path choice model. Vice-versa, the problem under study is that of calculating the O-D flows starting from the measured link flows, using network and path choice model (see Fig. 8.5.1).

Estimation of O-D matrices using traffic counts has received considerable attention in recent years both from the theoretical and the empirical point of view. This can be easily explained given the cost and complexity of sampling surveys, as well as the lack of precision related to both direct and model estimators of O-D flows. On the other hand, users' flows on some network links (traffic counts) are cheap and easily obtainable, often automatically. Furthermore, in many transportation-engineering applications, O-D flows estimates are essentially aimed at predicting traffic flows deriving from changes in the supply system (network). The focus is on estimating and predicting "aggregate" values of the O-D matrix, i.e. the traffic flows, rather than individual O-D flows and it is expected that a matrix capable of reproducing some of such aggregates with sufficient precision will give "better" predictions also following network changes.

Before solving the aggregate O-D estimation problem, it is necessary to express formally the relationship between the vector of observed flows and the unknown O-D demand flows, by reformulating some relationships presented in the previous chapters. As stated in Chapters 2 and 5, the flow f_l, in the reference period, can be expressed as the sum of flows on the paths including link l:

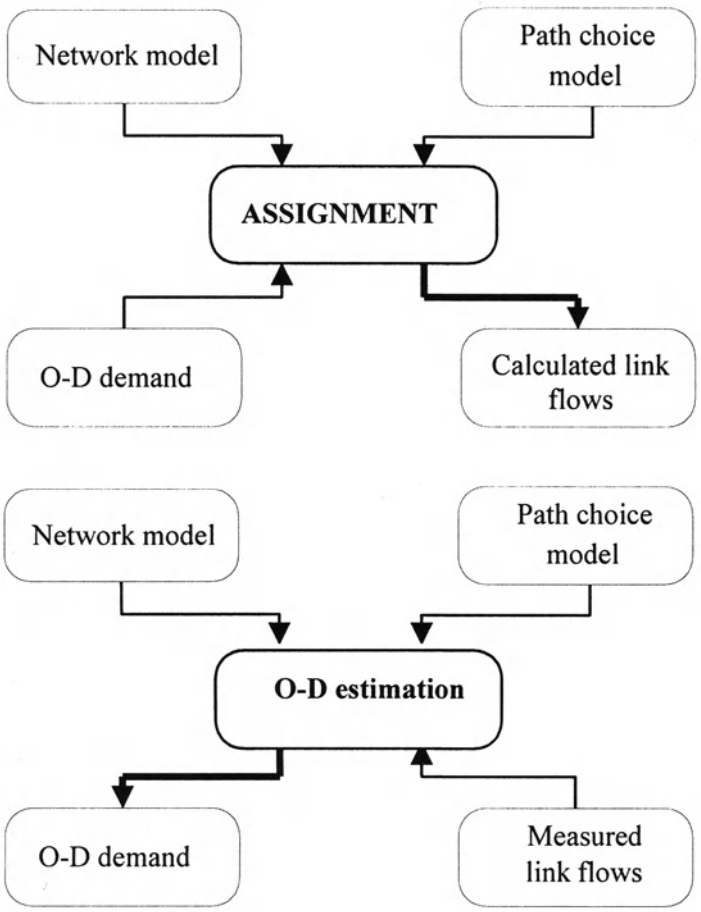


Fig. 8.5.1 Relationship between estimation of O-D flows with traffic counts and traffic assignment

$$f_l = \Sigma_k \, \delta_{lk} \, h_k$$

Path flows, can be expressed as the product of the O-D demand flow by the percentage (fraction) of users following each path:

$$f_l = \Sigma_k \, \delta_{lk} \, h_k = \Sigma_k \, \delta_{lk} \, \Sigma_i \, p_{ki} \, d_i \qquad (8.5.1)$$

where δ_{lk} is the element of the link-path incidence matrix Δ and p_{ki} is the fraction of

the flow $d_i^{(13)}$ between the i-th O-D pair using path k. Note that in the previous expression the variables (link flows, O-D flows, path flows and path fractions) indicate the "true" values relative to the system and to the reference period under study.

Equation (8.5.1) can be expressed differently as

$$f_l = \Sigma_i d_i \Sigma_k \delta_{lk} p_{ki} = \Sigma_i m_{li} d_i \tag{8.5.2}$$

or

$$f_l = \boldsymbol{m}_l^T \boldsymbol{d}$$

where $m_{li} = \Sigma_k \delta_{lk} p_{ki}$ is the assignment fraction, i.e. the fraction of the flow d_i using the link l and $\boldsymbol{m}_l$ is the column vector obtained by ordering these fractions[14]. Using a matrix notation, expression (8.5.2) becomes

$$\boldsymbol{f} = \boldsymbol{\Delta h} = \boldsymbol{\Delta P d} = \boldsymbol{M d} \tag{8.5.3}$$

All the variables introduced refer to the links for which traffic counts are available (n_l being their number), to the paths using them and to the O-D- flows using those paths (n_{OD} being their number). Thus, the matrix $\boldsymbol{M}$, or *assignment matrix*, has dimensions (n_l x n_{OD}). The relationship (8.5.3) between link flows and O-D demand flows is known as assignment relationship or map; Fig. 8.5.2 shows an example of the assignment map for an elementary network.

When several paths are available between each O-D pair, the elements of the assignment relationship, m_{li}, are not uniquely defined and therefore must be estimated. Path choice and network assignment models described in Chapters 4 and 5 provide methods for obtaining estimates $\hat{p}_{ki}$ of the fraction p_{ki} and estimates $\hat{m}_{li}$ of the fractions m_{li}.

In the case of pre-trip, deterministic or probabilistic path choice models, fractions $\hat{p}_{ki}$ can be expressed as probabilities of choosing each path k connecting the i-th O-D-pair as a function of the path costs vector $\boldsymbol{g}$ (see section 4.3.4.1):

$$\hat{p}_{ki} = p[k/i](\boldsymbol{g}) \tag{8.5.4}$$

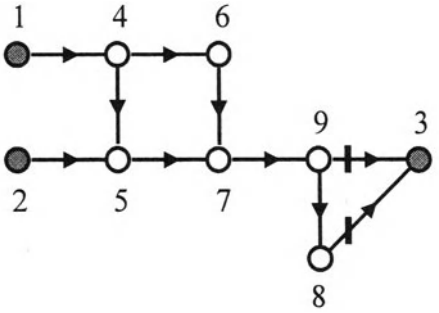
In the case of mixed pre-trip/en-route path choice models, usually adopted for high-frequency public transport networks, path probabilities can be obtained from hyperpath j choice probabilities $q[j/i]$, depending on the vector of hyperpath costs $\boldsymbol{x}$, and from the probabilities ω_{kj} of following path k within hyperpath j (see section 4.3.4.2):

$$\hat{p}_{ki} = \sum_j \omega_{kj} q[j/i](\boldsymbol{x}) \tag{8.5.5}$$

To underline the dependence of assignment matrix estimates $\hat{m}_{li}$ on the path choice model, and through this, on link costs c, the matrix $\hat{M}$ can be formally expressed as:

$$\hat{M} = \Delta \hat{\mathbf{P}}(c) \tag{8.5.6}$$

$$\hat{M} = \Delta \Omega Q(c) \tag{8.5.7}$$



$$d = \begin{bmatrix} d_{13} \\ d_{23} \end{bmatrix}$$

leaving as understood the relationship between additive path costs and link costs.

O-D pair	Path k							p_{ki}
1-3	1)	1	4	6	7	9	3	0.30
	2)	1	4	5	7	9	3	0.30
	3)	1	4	5	7	9	8 3	0.20
	4)	1	4	6	7	9	8 3	0.20
2-3	5)	2	5	7	9	3		0.70
	6)	2	5	7	9	8	3	0.30

$N = 2$ (*link* 9-3 *and link* 8-3)

Δ • P = M

	1	2	3	4	5	6
9-3	1	1			1	
8-3			1	1		1

•

	1-3	2-3
1	0.3	
2	0.3	
3	0.2	
4	0.2	
5		0.7
6		0.3

=

	1-3	2-3
9-3	0.6	0.7
8-3	0.4	0.3

Fig. 8.5.2 Assignment map for an elementary network.

If link and path costs are known[15], an estimate $\hat{\boldsymbol{M}}$ of the "true" assignment matrix $\boldsymbol{M}$ can be calculated through path choice models (8.5.4) and (8.5.5). It is to be expected that $\hat{\boldsymbol{M}}$ diverge from the true assignment matrix $\boldsymbol{M}$ because of the approximations implicit in any assignment model (network extraction, cost functions, path choice model, etc.). Thus, a vector ε^{SIM}[16] of assignment related errors should be added when substituting $\hat{\boldsymbol{M}}$ to $\boldsymbol{M}$ in equation (8.5.3):

$$\boldsymbol{f} = \boldsymbol{M}\boldsymbol{d} = (\hat{\boldsymbol{M}} + \boldsymbol{E}^{SIM})\,\boldsymbol{d} = \hat{\boldsymbol{M}}\,\boldsymbol{d} + \varepsilon^{SIM} \tag{8.5.8}$$

where $\boldsymbol{E}^{SIM}$ is the matrix of deviations between the true assignment matrix and that obtained with the assignment model and ε^{SIM} is the vector of deviations, or assignment errors, between the flows resulting from the assignment of "true" demand and "true" flows. In other words, if the "true" vector of demand flows $\boldsymbol{d}$[17] were known, its assignment to the network would produce a flows vector $\boldsymbol{v}$:

$$\boldsymbol{v} = \hat{\boldsymbol{M}}\,\boldsymbol{d} = \boldsymbol{v}(\boldsymbol{d}) \tag{8.5.9}$$

different from the "true" link flows vector $\boldsymbol{f}$. These deviations are the components of the vector ε^{SIM}:

$$\boldsymbol{f} = \boldsymbol{v} + \varepsilon^{SIM}$$

A further error source is related to flow counts. Like all measures, traffic counts are affected by errors depending, among other things, on the technique used (manual, automatic, etc.). Furthermore, the counts are usually conducted over several days, sometimes different for different network links, while the "true" demand vector $\boldsymbol{d}$ represents the average O-D flows in periods with similar characteristics (e.g. peak hour of the average weekday). Thus, if $\hat{\boldsymbol{f}}$ is the vector of measured flows, it will differ from the "true" vector $\boldsymbol{f}$ by a vector ε^{OBS} of measurement errors:

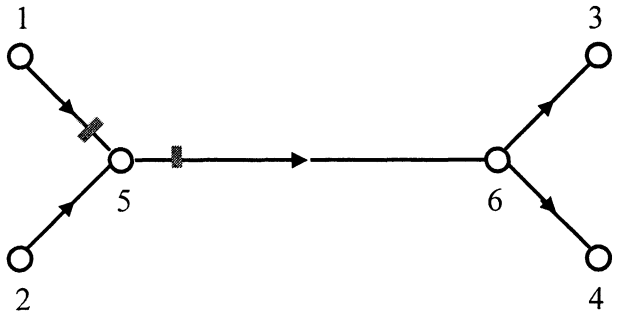
$$\hat{\boldsymbol{f}} = \boldsymbol{f} + \varepsilon^{OBS} \tag{8.5.10}$$

By combining the equation (8.5.8) and (8.5.10), it is possible to express the relationship between the vector of counts $\hat{\boldsymbol{f}}$, the assignment matrix $\hat{\boldsymbol{M}}$ and the "true" O-D demand flows vector $\boldsymbol{d}$ as:

$$\hat{\boldsymbol{f}} = \hat{\boldsymbol{M}}\,\boldsymbol{d} + \varepsilon^{SIM} + \varepsilon^{OBS} = \boldsymbol{v}(\boldsymbol{d}) + \varepsilon \tag{8.5.11}$$

where the vector ε is the algebraic sum of the vectors ε^{SIM} and ε^{OBS}. It is usually assumed that the assignment model and the counts are unbiased estimators of the "true" flows, i.e. that the vector ε is a zero mean random vector $E(\varepsilon)=\mathbf{0}$. Empirical evidence seems to support this assumption.

Usually information on O-D flows contained in traffic counts, represented by the system of stochastic equations (8.5.11), is not sufficient to estimate the vector $\boldsymbol{d}$. In fact, even assuming that ε is null, the independent equations in the linear system (8.5.3) are usually much less than the unknown O-D flows to be estimated. The example in Fig. 8.5.3 shows that even for an elementary network with a single path for each O-D pair, there are many O-D matrices which, once assigned to the network, can exactly reproduce the flows observed on the links.



$\hat{f}_{1,5} = 15 \quad \hat{f}_{5,6} = 20$

$\hat{d}$

		3	4
1		12	7
2		3	2

		3	4	$\sum_i (d_i - \hat{d}_i)^2$	$\frac{\sum_i (d_i - \hat{d}_i)^2}{\hat{d}_i}$
				Distance Measures	
d_1	1	5	10	66	8.702
	2	5	0		
d_2	1	7	8	26	2.226
	2	3	2		
d_3	1	12	3	16	2.286
	2	3	2		

Fig. 8.5.3 O-D matrices corresponding to the same link flows vector.

Furthermore, since in general the vector ε differs from zero, the system of linear equations $\hat{f} = \hat{M}d$ may not have a solution. In conclusion, the information contained in the counts must be combined with that from other sources to estimate the unknown O-D demand flows.

The additional information can be of two types: sampling or experimental information derived from demand surveys, and non-experimental information representing the a priori knowledge of the analyst. In the former case, reference can be made to the classic theory of statistical interference, while in the latter Bayesian estimators should be used. The two methods, whose statistical foundations will be described in the following sections, give rise to several estimators, some of them having similar formal representations[18].

In fact, if $\hat{d}$ is the vector representing the initial information, i.e. the information on O-D-demand not given by the flows, the ODCBE problem can be expressed in a general form as:

$$d^* = \underset{x \geq 0}{argmin}\left[z_1\left(x, \hat{d}\right) + z_2\left(v(x), \hat{f}\right)\right] \tag{8.5.12}$$

where x is the unknown demand vector. The two functions $z_1(x, \hat{d})$ and $z_2(v(x), \hat{f})$ can be considered respectively as "distance" measures of the unknown demand x from the a priori estimate $\hat{d}$ and of the flows obtained by assigning x to the network, $v(x)$, from the traffic counts $\hat{f}$. In an intuitive interpretation of the problem (8.5.12) is that of searching the vector d^* that is closest to the a priori estimate $\hat{d}$, and, once it is assigned to the network, produces the flows $v(d^*)$ closest to the counts $\hat{f}$.

In general, the functional form of the two terms z_1 (.) and z_2 (.), depends on the type of information available (experimental or non-experimental) and on the probability laws associated with such information. The statistical bases of the various estimators and their resulting functional forms will be described in the following sections.

8.5.1. Maximum Likelihood and GLS estimators*

Classic estimators of d can be specified following Maximum Likelihood theory or Generalized Least Squares (*GLS*) theory, depending on whether explicit assumptions on the probability distribution of random residuals ε^{SIM} and ε^{OBS} are made.

Maximum Likelihood (*ML*) estimators d^{ML} are obtained by maximizing the probability of observing sampling surveys results and counted flows. Under the usually acceptable assumption that these two probabilities are independent, the Maximum Likelihood estimator can be expressed as:

$$\boldsymbol{d}^{ML} = \underset{x \in S}{argmax} \left[lnL(\boldsymbol{n} / \boldsymbol{x}) + lnL(\hat{\boldsymbol{f}} / \boldsymbol{x}) \right] \tag{8.5.13}$$

where:

$\boldsymbol{x}$	is the "unknown" demand vector of dimensions ($n_{OD} \times 1$) whose components x_{od} are the trip flows between the O-D pair (o,d), from now on denoted with the double index;
$\boldsymbol{n}$	is the vector of demand counts with dimensions (n_{OD} x 1). The generic component of $\boldsymbol{n}$, n_{od}, is the number of trips between the O-D pair (o,d) observed in the sample;
$\hat{\boldsymbol{f}}$	is the vector of observed flows, or traffic counts, with dimension (n_l x 1).
$lnL(\boldsymbol{n}/\boldsymbol{x})$	is the log-likelihood function of demand counts, i.e. the logarithm of the probability of observing the sampling vector $\boldsymbol{n}$ if $\boldsymbol{x}$ is the (true) demand vector;
$lnL(\hat{\boldsymbol{f}}/\boldsymbol{x})$	is the log-likelihood function of the traffic counts, i.e. the logarithm of the probability of observing the vector of the counts $\hat{\boldsymbol{f}}$ if $\boldsymbol{x}$ is the (true) demand vector;
S	is the feasibility set of the (true) demand vector, usually coincident with the non-negative ortant, i.e. $S = \{\boldsymbol{x} : \boldsymbol{x} \geq 0\}$.

Maximum Likelihood estimators can be obtained therefore by solving the constrained maximization problem expressed by (8.5.13) once the log-likelihood functions $lnL(\boldsymbol{n}/\boldsymbol{x})$ and $lnL(\hat{\boldsymbol{f}}/\boldsymbol{x})$ have been specified. This requires the formulation of hypotheses on the probability laws of demand counts $\boldsymbol{n}$ and of traffic count $\hat{\boldsymbol{f}}$, conditional to the demand vector $\boldsymbol{x}$.

It is usually assumed that traffic counts are random variables with means given by the flows $\boldsymbol{v}(\boldsymbol{x})$ obtained by assigning the demand $\boldsymbol{x}$. This by (8.5.11) implies that the vector ε^{SIM} has a zero mean. The most widely used probability laws are Poisson and the multivariate normal. If it is assumed that the traffic counts on each link l are independently distributed as *Poisson* random variables with mean equal to $v_l(\boldsymbol{x})$, i.e.:

$$E[\hat{f}_l] = v_l(\boldsymbol{x}) = \hat{\boldsymbol{m}}_l^T \boldsymbol{x} \tag{8.5.14}$$

the probability of observing $\hat{\boldsymbol{f}}$ is given by the product of the probability of observing its individual components:

$$L(\hat{f}/x) = \prod_{l=1,\ldots,n_L} \frac{exp(-v_l(x))v_l(x)^{\hat{f}_l}}{\hat{f}_l!} \tag{8.5.15}$$

and the log-likelihood function becomes[19]:

$$lnL(\hat{f}/x) \cong \sum_{l=1..n_L} (\hat{f}_l \ln v_l(x) - v_l(x)) + const. \tag{8.5.16}$$

where the constant indicates other terms not depending on the unknown demand vector x and therefore irrelevant for the maximization problem (8.5.14).

If the traffic counts are jointly distributed according to a *Multivariate Normal* random variable with a mean vector equal to $v(x)$ and a variance-covariance matrix W the probability function of observing the vector $\hat{f}$ is proportional to:

$$L(\hat{f}/x) \propto exp[-\frac{1}{2}(\hat{f} - v(x))^T W^{-1}(\hat{f} - v(x))] \tag{8.5.17}$$

and the log-likelihood function becomes:

$$lnL(\hat{f}/x) = -\frac{1}{2}(\hat{f} - v(x))^T W^{-1}(\hat{f} - v(x)) + const. \tag{8.5.18}$$

The log-likelihood function of *demand counts* depends on the type of sampling adopted (see section 8.2). In the simplest case of stratified random sampling by zone of origin, a simple random sample of n_o trips is extracted from the d_o trips originating from each zone o (e.g. sampling at the entrances of a motorway network or at the cordon sections of the study area). Here it can be assumed that the number of trips sampled from each origin to all the destinations is distributed as a *multinomial* random variable. Further it can be assumed that the probability of observing the whole vector $\boldsymbol{n}$ is the product of the probability functions of these variables extended to all origins:

$$L(\boldsymbol{n}/x) = \prod_o \left[(n_o!/\prod_d n_{od}!) \prod_d (x_{od}/x_o)^{n_{od}} \right] \tag{8.5.19}$$

where x_{od}/x_o is the unknown probability of observing a trip with destination d.

From (8.5.19) the log-likelihood function can be obtained:

$$lnL(\boldsymbol{n}/x) = \Sigma_{od}\, n_{od} \ln x_{od} + const. \tag{8.5.20}$$

with the further constraint that the trips generated in each zone o are equal to those counted $d_{o.}$, or:

$$S=\{\boldsymbol{x} : \Sigma_d x_{od} = d_{o.}; \boldsymbol{x} \geq 0\}$$

If the number of trips sampled at each origin is sufficiently large (a few dozen), the multinomial variable can be closely approximated by the product of independent Poisson variables (one for each O-D pair), with parameters equal to the means $\alpha_o x_{od}$, where α_o is the sampling rate in origin o:

$$\alpha_o = \frac{n_o}{d_{o.}}$$

In this case the functions $L(\boldsymbol{n}/\boldsymbol{x})$ and $lnL(\boldsymbol{n}/\boldsymbol{x})$ given by (8.5.19) and (8.5.20) can be approximated by:

$$L(\boldsymbol{n}/\boldsymbol{x}) = \Pi_{od} \frac{exp(-\alpha_o x_{od})(\alpha_o x_{od})^{n_{od}}}{n_{od}!} \tag{8.5.21}$$

and

$$ln\, L(\boldsymbol{n}/\boldsymbol{x}) = \Sigma_{od} (n_{od}\, ln(\alpha_o x_{od}) - \alpha_o x_{od}) + const. \tag{8.5.22}$$

Analogous expressions can be obtained for more complex sampling methods; in applications, however, expressions (8.5.20) and (8.5.22) are often used as reasonable approximations.

In conclusion, the Maximum Likelihood estimator $\boldsymbol{d}^{ML}$ is obtained by substituting expression (8.5.16), or (8.5.18) for the log-likelihood function of traffic counts and expression (8.5.20) or (8.5.22) for demand counts in the general expression (8.5.13) (see fig. 8.5.4).

Generalized Least Squares (*GLS*) is the other estimator derived within the classic theory of statistical inference. The *GLS* estimator provides the estimate of an unknown vector, in this case the demand flows vector, starting from a system of linear stochastic equations. The latter can be obtained by combining the information on demand contained in the traffic counts, expressed by the equation (8.5.11), and in the direct estimate $\hat{d}$, obtained from demand counts:

$$\begin{aligned} \hat{\boldsymbol{f}} &= \hat{\boldsymbol{M}}\boldsymbol{x} + \boldsymbol{\varepsilon} \qquad E(\boldsymbol{\varepsilon}) = \boldsymbol{0}\ Var[\boldsymbol{\varepsilon}] = \boldsymbol{W} \\ \hat{\boldsymbol{d}} &= \boldsymbol{x} + \boldsymbol{\eta} \qquad E(\boldsymbol{\eta}) = \boldsymbol{0}\ Var[\boldsymbol{\eta}] = \boldsymbol{Z} \end{aligned} \tag{8.5.23}$$

where $\hat{\boldsymbol{d}}$ is the O-D demand vector whose components $\hat{d}_{od}$ are the sample estimates, obtained with the methods described in section 8.2. For example, in the case of simple random sampling with rate α, these estimates will be:

$$\hat{d}_{od} = \frac{n_{od}}{\alpha}$$

The vector η in expression (8.5.23) is the vector of sampling errors whose components are the deviations between the true unknown demand $\boldsymbol{x}$ and the sample estimates $\hat{\boldsymbol{d}}$. If the estimator adopted is unbiased the vector η has zero mean. The elements of the variance-covariance matrix $\boldsymbol{Z}$ can be estimated by using the relevant expressions for variances and covariances of sample estimates.

The *GLS* estimator of the demand vector can therefore be expressed as:

$$\boldsymbol{d}^{GLS} = \underset{x \in S}{arg\,min}\left[(\hat{\boldsymbol{d}} - \boldsymbol{x})^T \boldsymbol{Z}^{-1}(\hat{\boldsymbol{d}} - \boldsymbol{x}) + (\hat{\boldsymbol{f}} - \hat{\boldsymbol{M}}\boldsymbol{x})^T \boldsymbol{W}^{-1}(\hat{\boldsymbol{f}} - \hat{\boldsymbol{M}}\boldsymbol{x})\right] \qquad (8.5.24)$$

Expression (8.5.24) is often applied, assuming that the matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ are diagonal, i.e. ignoring the covariances between the components of vectors ε^{SIM} and η, both because these are difficult to express and memory occupation and computing times should be reduced. Under this simplified assumption, expression (8.5.24) becomes:

$$\boldsymbol{d}^{GLS} = \underset{x \geq 0}{arg\,min}\left[\sum_{od} \frac{(\hat{d}_{od} - x_{od})^2}{Var[\eta_{od}]} + \sum_{l} \frac{(\hat{f}_l - \sum_{od} \hat{m}_{l,od} x_{od})^2}{Var[\varepsilon_l]}\right] \qquad (8.5.25)$$

The intuitive interpretation given for (8.5.12) can be extended to (8.5.25): the demand vector estimated using the counted flows, $\boldsymbol{d}^{GLS}$, minimizes the sum of quadratic deviations compared to the initial sampling estimate and of assigned flows with the counted flows. Furthermore, the quadratic deviations have weights inversely proportional to the variances of their respective errors. In other words, deviation from a component $\hat{d}_{od}$ will weigh less the "worse" is the estimate, i.e. the greater is the $Var\,[\eta_{rs}]$ and the same is true for the flows.

Note also the role of information on the vector $\boldsymbol{d}$ contained in traffic counts. If this information did not exist, the second term of equations (8.5.24) and (8.5.25) would disappear and the estimate of $\boldsymbol{d}^{GLS}$ would coincide with $\hat{\boldsymbol{d}}$ since the latter minimizes the quadratic objective function setting it to zero. Similar considerations can be made for the Maximum Likelihood estimators.

From the formal point of view, the *GLS* estimator coincides with the Maximum Likelihood estimators if both the demand estimates and traffic counts are assumed to be distributed as multivariate normal random variables with mean equal to $\boldsymbol{x}$ and $\hat{\boldsymbol{M}}\boldsymbol{x}$ and dispersion matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ respectively.

8.5.2. Bayesian estimators*

Bayesian methods estimate unknown parameters by combining experimental, or sampling, information with non-experimental, or "subjective", information[20]. In the particular case of O-D demand estimation, experimental information is relative to traffic counts, while non-experimental information may be relative to old O-D estimates to be updated, to estimates obtained with demand models or simply to the analyst "expectations". In each case, $\hat{\boldsymbol{d}}$ will indicate the demand vector composed of non-experimental estimates. Bayesian estimators are obtained from the a posteriori probability function $h(\boldsymbol{x}/\hat{\boldsymbol{f}}, \hat{\boldsymbol{d}})$ of the unknown demand vector $\boldsymbol{x}$ conditional on a priori information $\hat{\boldsymbol{d}}$ and on experimental information $\hat{\boldsymbol{f}}$. According to Bayes theorem, the a posteriori probability is proportional to the product of the a priori probability function $g(\boldsymbol{x}/\hat{\boldsymbol{d}})$, expressing the distribution of subjective probability attributed to the unknown vector given the a priori estimate $\hat{\boldsymbol{d}}$, and the probability, or Likelihood, function $L(\hat{\boldsymbol{f}}/\boldsymbol{x})$ expressing the probability of observing the traffic counts $\hat{\boldsymbol{f}}$ conditional to the unknown demand vector $\boldsymbol{x}$:

$$h(\boldsymbol{x}/\hat{\boldsymbol{f}}, \hat{\boldsymbol{d}}) \propto L(\hat{\boldsymbol{f}}/\boldsymbol{x})\, g(\boldsymbol{x}/\hat{\boldsymbol{d}}) \tag{8.5.26}$$

A family of Bayesian estimators for demand flows, $\boldsymbol{d}^B$, can be obtained by maximizing[21] the a posteriori probability (8.5.26) or its natural logarithm:

$$\boldsymbol{d}^B = \underset{\boldsymbol{x}\in S}{argmax}\left[ln\, g(\boldsymbol{x}/\hat{\mathbf{d}}) + lnL(\hat{\mathbf{f}}/\boldsymbol{x}) \right] \tag{8.5.27}$$

The specification of Bayesian estimators depends on the assumptions made for the probability functions $L(\hat{\boldsymbol{f}}/\boldsymbol{x})$ and $g(\boldsymbol{x}/\hat{\boldsymbol{d}})$. With respect to the function $L(\hat{\boldsymbol{f}}/\boldsymbol{x})$, equations (8.5.16) and (8.5.18), corresponding to the assumptions of independent Poisson and multivariate normal random variables respectively, can be used.

The a priori probability function, $g(\boldsymbol{x}/\hat{\boldsymbol{d}})$, can be specified in different ways; the formulations proposed in literature are described below.

If it is assumed that the unknown demand vector is a *multinomial random variable* resulting from the distribution of total demand d among all possible O-D pairs, with probabilities π_{od} given by the matrix $\hat{\boldsymbol{d}}$:

$$\pi_{od} = \frac{\hat{d}_{od}}{\hat{d}..} \qquad \hat{d}.. = \sum_{od} \hat{d}_{od}$$

the function $g(\boldsymbol{x}/\hat{\boldsymbol{d}})$ can be written as:

$$g\left(\boldsymbol{x}/\hat{\boldsymbol{d}}\right) = \frac{\left(\sum_{od} x_{od}\right)!}{\Pi_{od} x_{od}!} \Pi_{od} \left(\hat{d}_{od} / \hat{d}..\right)^{x_{od}} \tag{8.5.28}$$

Using the Stirling approximation ($ln\, x! \cong x\, ln\, x\text{-}x$), the logarithm of (8.5.28) can be expressed as:

$$\ln g\left(\boldsymbol{x}/\hat{\boldsymbol{d}}\right) = \Sigma_{od}\ x_{od}\ \ln\left(\Sigma_{od}\, x_{od}\right) + \Sigma_{od}\ x_{od}\ \ln\left(\hat{d}_{od}/\hat{d}..\, x_{od}\right) \tag{8.5.29}$$

Furthermore, if the total number of trips $\left(\Sigma_{od}\, x_{od} = \hat{d}..\right)$ is assumed to be known, expression (8.5.29) is further simplified in:

$$\ln\ g\left(\boldsymbol{x}/\hat{\boldsymbol{d}}\right) = -\Sigma_{od}\, x_{od}\, \ln\left(x_{od} / \hat{d}_{od}\right) + const. \tag{8.5.30}$$

The opposite of function (8.5.30) is known in literature as the *entropy function* of the unknown vector $\boldsymbol{x}$.

Alternatively, it can be assumed that the components x_{od} are independently distributed as *Poisson random variables*, with mean (parameter) equal to $\hat{d}_{od}$. In this case the function $g(\boldsymbol{x}/\hat{\boldsymbol{d}})$ becomes:

$$g\left(\boldsymbol{x}/\hat{\boldsymbol{d}}\right) = \Pi_{od} \frac{\exp\left(-\hat{d}_{od}\right)^{x_{od}}}{x_{od}!} \hat{d}_{od}^{x_{od}} \tag{8.5.31}$$

The latter, using Stirling's approximation, can be expressed as:

$$\ln g\left(\boldsymbol{x}/\hat{\boldsymbol{d}}\right) = -\Sigma_{od}\, x_{od} \left[\ln\left(x_{od} / \hat{d}_{od}\right) - 1\right] + const. \tag{8.5.32}$$

The opposite of function (8.5.32) is also known in the literature as *information function* of the unknown vector $\boldsymbol{x}$.

Finally, it can be assumed that the vector $\boldsymbol{x}$ is distributed according to a *Multivariate Normal random variable* of mean $\hat{\boldsymbol{d}}$ and variance-covariance matrix $\boldsymbol{Z}_B$; in this case the probability function is proportional to:

$$g\left(x/\hat{d}\right) \propto \exp\left[-\frac{1}{2}\left(x-\hat{d}\right)^{T} Z_B^{-1}\left(x-\hat{d}\right)\right]$$

and its logarithm becomes:

$$ln\, g\left(x/\hat{d}\right) = -\frac{1}{2}\left(x-\hat{d}\right)^{T} Z_B^{-1}\left(x-\hat{d}\right) + const. \qquad (8.5.33)$$

If the a priori probability function $g(x/\hat{d})$ and the traffic counts probability function $L(\hat{f}/x)$ are both assumed to be multivariate normal variables, expressions (8.5.18) and (8.5.33) are substituted in the general expression (8.5.27) and the resulting Bayesian estimator is formally analogous to the Generalized Least Squares estimator d^{GLS}. However, the similarity between the two estimators is only formal, since the vector $\hat{d}$ and the variance-covariance matrices Z and Z_B have different interpretations. In the *GLS* estimator, the vector $\hat{d}$ is a direct demand estimate from sampling surveys and the matrix Z includes its sampling variances and covariances. In Bayesian estimators $\hat{d}$ is an a priori estimate of the O-D demand vector and Z_B is made up of variances and covariances summarizing the analyst confidence in such estimate.

The formal analogy of the two estimators should, however, be considered an advantage since it allows the use of the same model and algorithm in very different estimation situations. This generality of *GLS* estimator has contributed to its widespread use in applications.

8.5.3. Applicative issues

We stated that different estimators combining traffic counts $\hat{f}$ and other information $\hat{d}$ can be expressed in a general form as the vector d^* solving a constrained minimization problem[22]:

$$d^* = \underset{x \in S}{arg\,min}\left[z_1\left(x,\hat{d}\right) + z_2\left(v(x),\hat{f}\right)\right] \qquad (8.5.34)$$

Fig. 8.5.4 summarizes the functional forms of $z_1(\cdot)$ and $z_2(\cdot)$ previously described and the corresponding assumptions.

The application of these methods in practice poses several problems briefly addressed below.

GENERAL ESTIMATION MODEL

$$d^* = \underset{x \in S}{arg\,min}\left[z_1\left(x, \hat{d}\right) + z_2\left(v(x), \hat{f}\right)\right]$$

Distance from the initial estimate $z_1(x, \hat{d})$	Distance from flow counts $z_2(v(x), \hat{f})$
Generalized Least Squares (GLS) $(\hat{d} - x)^T Z^{-1} (\hat{d} - x)$ or $\sum_{od} (x_{od} - \hat{d}_{od})^2 / Var[\eta_{od}]$	Generalized Least Squares (GLS) $(\hat{f} - v(x))^T W^{-1} (\hat{f} - v(x))$ or $\sum_{l \in M} (\hat{f}_l - v_l(x))^2 / Var[\varepsilon_l]$
Maximum Likelihood (ML) Poisson $-\sum_{od} (n_{od} \ln(\alpha_{ods} x_{od}) - \alpha_{od} x_{od})$ Multinomial $-\sum_{od} n_{od} \ln x_{od}$	Maximum Likelihood (ML) Poisson $-\sum_{l \in M} (\hat{f}_l \ln v_l(x) - v_l(x))$ MVN $(\hat{f} - v(x))^T W^{-1} (\hat{f} - v(x))$ or $\sum_{l \in M} (\hat{f}_l - v_l(x))^2 / Var[\varepsilon_l]$
Bayes Poisson $\sum_{od} x_{od} ln[(x_{od} / \hat{d}_{od}) - 1]$ MVN $(\hat{d} - x)^T Z^{-1} (\hat{d} - x)$ or $\sum_{od} (x_{od} - \hat{d}_{od})^2 / Var[\eta_{od}]$ Multinomial $\sum_{od} x_{od} ln(x_{od} / \hat{d}_{od})$	Bayes Poisson $-\sum_{l \in M} (\hat{f}_l \ln v_l(x) - v_l(x))$ MVN $(\hat{f} - v(x))^T W^{-1} (\hat{f} - v(x))$ or $\sum_{l \in M} (\hat{f}_l - v_l(x))^2 / Var[\varepsilon_l]$

Fig. 8.5.4 Functional forms of the terms of $z_1(\cdot)$ and $z_2(\cdot)$.

The *choice of functional form* from among the various possibilities obviously depends on the type of available information about the O-D flows and therefore on the estimation context (classic or Bayesian). The Generalized Least Squares estimator is "robust" since it can be adopted in both cases and, as a classic estimator, does not require explicit assumptions on the probability, or likelihood, law of traffic and demand counts. Obviously this "robustness" is paid for in terms of statistical properties which are less satisfactory than those of other estimators if probability distributions are known for traffic and demand counts.

The literature presents a number of simulation studies comparing the statistical performances of various estimators. Statistical performances can be measured by the "divergence" between the estimates $\boldsymbol{d}^*$ obtained for different specifications of the model (8.5.34) and the true demand vector $\boldsymbol{d}$ used in the simulation. The Mean Square Error between the two demand vectors, $MSE(\boldsymbol{d}^*,\boldsymbol{d})$, is one of the most popular divergence measures:

$$MSE(\boldsymbol{d}^*,\boldsymbol{d}) = \frac{1}{n_{OD}} \Sigma_{od} \left(d^*_{od} - d_{od}\right)^2$$

where n_{OD} is the number of O-D pairs.

Alternatively, the ratio between the square root of the Mmean Square Error and the average demand, analogous to the coefficient of variation of a random variable, can be adopted:

$$RMSE\% = \frac{MSE(\boldsymbol{d}^*,\boldsymbol{d})^{1/2}}{d../n_{OD}}$$

Obviously, the estimator $\boldsymbol{d}^*$ is the better the lower the *MSE* and *RMSE%* are. Numerical results seem to confirm the theoretical indications and suggest that the *GLS* estimator gives more stable results compared with other estimators under a wide range of hypotheses on the information contained in $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{f}}$.

The use of *GLS* estimators requires the *definition of variance-covariance matrices* $\boldsymbol{Z}$ and $\boldsymbol{W}$. This issue arises only in the case of *GLS* estimators and should be seen as a further degree of freedom since variances and covariances are implicitly defined by the distributions underlying the other functional forms of $z_1(\cdot)$ and $z_2(\cdot)$. For example, expression (8.5.16) for $z_2(\boldsymbol{v}(\boldsymbol{x}), \hat{\boldsymbol{f}})$ implies the assumption that traffic counts are independent Poisson variables, their deviations from the mean $\boldsymbol{v}(\boldsymbol{x})$ are independent ($Cov(\varepsilon_l\varepsilon_m)=0$) and their variance is equal to the mean ($Var[\varepsilon_l]=v_l(x)$).

In applications, covariances among the components of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ are usually ignored, i.e. matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ are assumed to be diagonal. If $\hat{\boldsymbol{d}}$ is a sample estimate the variance of sampling error η_{od} depends on the sampling strategy and can be computed, e.g. by using the formulae (8.2.3) and (8.2.8). In Bayesian estimation

variances are a measure of the analyst "confidence" in the a priori estimates and therefore cannot be univocally defined. The variances of residuals ε_l^{OBS} can be obtained through empirical relationships expressing the coefficient of variation CV of assignment errors for different assignment models as a function of measured flows. An example of this type of results was shown in Fig. 5.8.2 of Chapter 5.

8.5.4. Solution methods

The main computational problem in solving the ODCBE models is the *calculation of the assignment map* $v(x)$, i.e. the *assignment matrix* $\hat{\boldsymbol{M}}$ expressed by (8.5.6) and (8.5.7). The elements $\hat{m}_{li}$, depend on path choice probabilities, (8.5.4) and (8.5.5), which in turn are functions of path (or hyperpath) costs and, thus, of link costs as formally expressed by equations (8.5.6) and (8.5.7).

In general, given the path or hyperpath choice model, computation of the assignment matrix for given costs c can be conducted with relatively straightforward modifications to the network loading algorithms described in section 7.3. Furthermore, in the case of congested networks, link costs depend on the link flows vector $\boldsymbol{f}$. Solution of the ODCBE problem has two levels of complexity according to whether the link costs vector is known or not.

Link costs known. Let us assume that an estimate $\hat{c}$ of link cost is available. This is the case if the network is uncongested, or moderately congested, and link costs can be estimated independently of the flows. Alternatively, link costs can be estimated for congested networks either directly through network (travel times) surveys or indirectly through cost functions and flow counts on congested links $\hat{c}_l = c_l(\hat{f}_l)$. Direct network surveys can be carried out automatically with surveillance systems based on vehicle location and remote transmission through TLC technologies.

If link costs are known, it is possible to estimate the assignment matrix $\hat{\boldsymbol{M}}(\hat{c})$ independently of the demand vector; thus $\boldsymbol{d}^*$ can be estimated by applying the model (8.5.34):

$$\boldsymbol{d}^* = \arg\min_{x \in S} \left[z_1(\boldsymbol{x}, \hat{\boldsymbol{d}}) + z_2(\hat{\boldsymbol{M}}(\hat{c})\boldsymbol{x}, \hat{\boldsymbol{f}}) \right] \tag{8.5.35}$$

Model (8.5.35) is a constrained minimization problem that can be solved with different algorithms, depending on the constraints defining the set S. Often the feasibility set S is defined by non-negativity constraints for the demand flows ($x_i \geq 0$ $\forall i$). The projected gradient algorithm, described in Appendix A, can be used in this case. It is usually possible to formulate explicitly the gradient of the objective function. For the *GLS* estimator, under the assumption that the matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ are diagonal, the i-th component of the gradient can be expressed as:

$$Gr_i = \frac{\partial}{\partial x_i}\left[\sum_i \frac{\left(x_i - \hat{d}_i\right)^2}{Var[\eta_i]} + \sum_l \frac{\left(\hat{f}_l - \sum_i \hat{m}_{li} x_i\right)^2}{Var[\varepsilon_l]}\right] =$$

$$= 2\left[\frac{\left(x_i - \hat{d}_i\right)}{Var[\eta_i]} + \sum_l \frac{\hat{m}_{li}\left(\sum_j \hat{m}_{lj} x_j - \hat{f}_l\right)}{Var[\varepsilon_l]}\right]$$

Fig. 8.5.5 reports the main variables of an application of the projected gradient algorithm for the calculation of $\boldsymbol{d}^{GLS}$ on a test network.

Link costs unknown. Estimates of link costs might not be available for all links. This is typically the case with congested networks, when the information described above is not available. In this case a problem of circular dependence arises, since it is possible to estimate link flows $\boldsymbol{v}(\boldsymbol{d}^*)$, and therefore link costs $\boldsymbol{c}(\boldsymbol{v}(\boldsymbol{d}^*))$, by assigning the demand $\boldsymbol{d}^*$, solution to the problem (8.5.35), which in its turn is estimated from link flows and costs. The estimation problem can be formalized as a fixed-point problem as described below.

Let $\boldsymbol{d}=\delta(\hat{\boldsymbol{M}})$ be the solution of the estimation problem (8.5.35) for a given assignment matrix $\hat{\boldsymbol{M}}$:

$$\boldsymbol{d} = \delta\left(\hat{\boldsymbol{M}}\right) = \underset{\boldsymbol{x}\in S}{arg\,min}\left[z_1(\boldsymbol{x}, \hat{\boldsymbol{d}}) + z_2(\hat{\boldsymbol{M}}\boldsymbol{x}, \hat{\boldsymbol{f}})\right]$$

If the above problem has only one solution, the relationship $\boldsymbol{d} = \delta(\hat{\boldsymbol{M}})$ can be considered to be a function associating to each assignment matrix $\hat{\boldsymbol{M}}$ an estimate of the demand vector $\boldsymbol{d}$. The assignment matrix $\hat{\boldsymbol{M}}$ can be expressed as a function of demand flows. In fact, if we combine the relationship connecting the assignment matrix to link costs, $\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}(\boldsymbol{c})$, with the cost functions $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{f})$, and introduce the relationship between link and demand flows through the assignment model, $\boldsymbol{f} = \boldsymbol{v}(\boldsymbol{d})$, we get:

$$\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}\,(\boldsymbol{c}(\boldsymbol{v}(\boldsymbol{d})))$$

Thus the ODCBE problem can be expressed through a fixed-point model obtained by combining the two functions $\boldsymbol{d}=\delta(\hat{\boldsymbol{M}})$ and $\hat{\boldsymbol{M}} = \hat{\boldsymbol{M}}\,(\boldsymbol{c}(\boldsymbol{v}(\boldsymbol{d})))$:

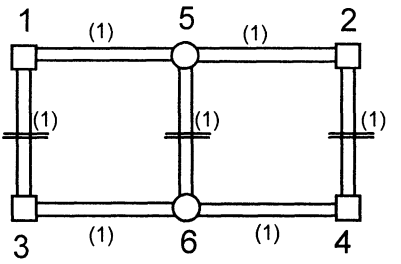
$$\boldsymbol{d}^* = \delta(\hat{\boldsymbol{M}}\,(\boldsymbol{d}^*))$$

or

$$\boldsymbol{d}^* = \underset{\boldsymbol{x}\in S}{argmin}\left[z_1(\boldsymbol{x}, \hat{\boldsymbol{d}}) + z_2(\hat{\boldsymbol{M}}(\boldsymbol{c}(\boldsymbol{v}(\boldsymbol{d}^*)))\boldsymbol{x}, \hat{\boldsymbol{f}})\right] \qquad (8.5.36)$$

TRUE O/D MATRIX

-	25	25	25
25	-	25	25
25	25	-	25
25	25	25	-



(.) link costs $\hat{c}_l$

TRUE DEMAND VECTOR

$$\boldsymbol{d}^T = [25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25 \quad 25]$$

INITIAL DEMAND VECTOR

$$\hat{\boldsymbol{d}}^T = [0 \quad 50 \quad 0 \quad 0 \quad 50 \quad 50 \quad 50 \quad 0 \quad 50 \quad 0 \quad 100 \quad 0]$$

COUNTED FLOWS

$$\hat{\boldsymbol{f}}^T = [49 \quad 49 \quad 33 \quad 33 \quad 49 \quad 49]$$

PATH CHOICE MODEL

$$\hat{p}_{ki} = \frac{exp(-C_i)}{\sum_k exp(-C_k)}$$

ASSIGNMENT MATRIX *M*

O/D pairs		1	2	3	4	5	6	7	8	9	10	11	12
links		1-2	1-3	1-4	2-1	2-3	2-4	3-1	3-2	3-4	4-1	4-2	4-3
1	1-3	20	86	33		33	2						20
2	3-1				20			86	33	20	33	2	
3	5-6	10	12	33	10	33	12			10			10
4	6-5	10			10			12	33	10	33	12	10
5	2-4		2	33	20	33	86			20			
6	4-2	20						2	33		33	86	20

ASSIGNED FLOWS

Fig. 8.5.6a Application of the projected gradient algorithm for the computation of d^{GLS} (input data).

	$z^k()$	|h|		1	2	3	4	5	6	7	8	9	10	11	12
	425	3.658	d	0	50	0	0	50	50	50	0	50	0	100	0
1			$-\nabla$	-0.905	-1.195	-1.089	-0.254	-1.089	-1.352	-0.081	-0.823	-0.254	-0.823	-2.978	-0.905
			h	0	-1.195	0	0	-1.089	-1.352	-0.084	0	-0.254	0	-2.978	0
	92	1.347	d	0	35	0	0	36	33	49	0	47	0	63	0
2			$-\nabla$	0.194	0.340	0.502	0.379	0.517	0.347	0.248	0.444	0.383	0.444	-0.487	0.194
			h	0.194	0.340	0.502	0.379	0.517	0.347	0.248	0.444	0.383	0.444	-0.487	0.194
	62	0.817	d	2	38	5	4	41	36	51	4	50	4	58	2
3			$-\nabla$	-0.078	-0.282	-0.215	-0.018	-0.201	-0.329	-0.281	0.036	-0.015	0.056	-0.544	-0.078
			h	-0.078	-0.282	-0.215	-0.018	-0.201	-0.329	-0.281	0.056	-0.015	0.056	-0.544	-0.078
	47	0.647	d	1	35	2	3	39	33	48	5	50	5	53	1
4			$-\nabla$	0.145	0.108	0.182	0.162	0.206	0.072	-0.077	0.314	0.166	0.314	-0.151	0.145
			h	0.145	0.108	0.182	0.162	0.206	0.072	-0.077	0.314	0.166	0.314	-0.151	0.145
	9.92	0.159	d	5	33	4	9	43	29	33	19	56	19	37	5
9			$-\nabla$	-0.025	-0.049	-0.054	-0.020	-0.042	-0.053	-0.082	-0.020	-0.017	-0.020	-0.078	-0.025
			h	-0.025	-0.049	-0.054	-0.020	-0.042	-0.053	-0.082	-0.020	-0.017	-0.020	-0.078	-0.025
	9.45	0.099	d	5	32	4	8	42	29	33	18	56	18	36	5
10			$-\nabla$	0.019	0.021	0.021	0.025	0.033	0.018	0.008	0.052	0.028	0.052	-0.005	0.019
			h	0.019	0.021	0.021	0.025	0.033	0.018	0.008	0.052	0.028	0.052	-0.005	0.019

	1	2	3	4	5	6	7	8	9	10	11	12
True O/D vector	25	25	25	25	25	25	25	25	25	25	25	25
Initial O/D vector	0	50	0	0	50	50	50	0	50	0	100	0
Estimated O/D vector	5	32	4	8	42	29	33	18	56	18	36	5

STATISTICS

MSE (true – estimate)	3120.894
MSE (true – initial)	12500.0
Percentage reduction of MSE	0.7503

Fig. 8.5.6b Application of the projected gradient algorithm for the computation of $\mathbf{d}^{GLS}$ (main variables and comparison statistics).

Alternatively, the ODCBE problem for congested networks can be stated as a *bi-level optimization problem*. This is the case when the equilibrium assignment map is expressed through an optimization model, as described in section 5.4 and 5.A for DUE and SUE respectively. In this case the problem can be stated formally as:

$$\boldsymbol{d}^* = \underset{\boldsymbol{x}\in S}{argmin}\left[z_1(\boldsymbol{x},\hat{\boldsymbol{d}}) + z_2(\boldsymbol{v}(\boldsymbol{x}),\hat{\boldsymbol{f}})\right] \tag{8.5.37}$$

$$\boldsymbol{v}(\boldsymbol{x}) = \underset{\boldsymbol{f}\in S_f(\boldsymbol{x})}{argmin}\, z(\hat{\boldsymbol{f}})$$

where $z(\cdot)$ is the objective function corresponding to the DUE or SUE equivalent optimization problem and the dependence of the link flows feasibility set on the demand vector has been stated explicitly. Obviously the bi-level optimization approach requires that the assignment problem can be expressed by an optimization model, i.e. it satisfies the mathematical properties stated in Chapter 5, such as continuous cost functions with symmetric Jacobian. If this is the case, the two formulations (8.5.36) and (8.5.37) are equivalent.

Problems (8.5.36) and (8.5.37) are computationally more complex than problem (8.5.35) since it is necessary to simultaneously solve the constrained optimization problem (8.5.35) given the demand estimate, and the equilibrium assignment problem yielding the link flows and costs[23]. The fixed-point problem (8.5.36) can be solved by using fixed-point iterative algorithms, which alternately solve the demand estimation and assignment problems by averaging out the results until convergence. For example, the MSA algorithm described in Appendix A and applied in Chapter 7 to calculate SUE equilibrium flows can be adopted. The structure of the algorithm can be represented for the generic iteration k, assuming that a current estimate $\boldsymbol{d}^{k-1}$ is available from the previous iteration. The main steps are as follows:

- Calculation of assignment matrix $\hat{\boldsymbol{M}}^k$ corresponding to demand $\boldsymbol{d}^{k-1}$
  - Assignment of demand $\boldsymbol{d}^{k-1}$ to the network and computation of the corresponding flows

$$\boldsymbol{v}^k = \boldsymbol{v}(\boldsymbol{d}^{k-1})$$

  - Estimation of link costs and assignment matrix with the obtained flows

$$\boldsymbol{c}^k = \boldsymbol{c}(\boldsymbol{v}^k)$$
$$\hat{\boldsymbol{M}}^k = \hat{\boldsymbol{M}}(\boldsymbol{c}^k)$$

- Estimation of demand support vector $\boldsymbol{y}^k$

$$\boldsymbol{y}^k = \underset{x \in S}{arg\,min}[z_1(\boldsymbol{x}, \hat{\boldsymbol{d}}) + z_2(\hat{\boldsymbol{M}}^k \boldsymbol{x}, \hat{\boldsymbol{f}})]$$

- Updating the demand estimate with a "weighted average" of $\boldsymbol{d}^{k-1}$ and $\boldsymbol{y}^k$:

$$\boldsymbol{d}^k = \frac{k-1}{k}\boldsymbol{d}^{k-1} + \frac{1}{k}\boldsymbol{y}^k$$

This procedure is repeated until a suitable termination test ($\boldsymbol{y}^k \cong \boldsymbol{d}^{k-1}$) is satisfied. The MSA algorithm could be applied to other variables such as link costs or assignment fractions.

8.6. Aggregate calibration of demand models using traffic counts

The aggregate information on transportation demand contained in traffic counts[24] can also be used to estimate the parameters (calibration) of demand models. As

stated in Chapter 4, demand models can be seen as functions relating the demand flows to variables of the activities system, ***SE***, and of the transport system, ***T***, through a vector of unknown parameters β[25].

$$\boldsymbol{d} = \boldsymbol{d}(\boldsymbol{SE}, \boldsymbol{T}; \beta) \tag{8.6.1}$$

For a given specification of the model and given values of ***SE*** and ***T***, expression (8.6.1) can be considered to be a relationship between demand flows and the unknown vector β. This section discusses the problem of combining traffic counts with other information (experimental or not) to estimate unknown parameters β. As in the case of O-D- flow estimation, the problem can be formulated following classic and Bayesian approaches.

The first case arises when other experimental information is available, typically RP or SP sampling surveys, for the calibration of demand models. The estimates $\hat{\beta}$ resulting from the methods described in sections 8.3 and 8.4 can be seen as determinations of random variables; thus the estimate of the generic component, $\hat{\beta}_i$, diverges from the "real" value by an unknown quantity σ_i:

$$\hat{\beta}_i = \beta_i + \sigma_i \tag{8.6.2}$$

If $\hat{\beta}_i$ is a Maximum Likelihood estimate, the variance of σ_i can be calculated by the inverse of the Hessian matrix of the Log-Likelihood function, see equation (8.3.6). Furthermore if the estimator is (asymptotically) unbiased $E(\sigma_i) \rightarrow 0$.

Alternatively, in a Bayesian approach, $\hat{\beta}$ can include a priori expectations on the parameters, e.g. values obtained in a similar study area. In this case, expression (8.6.2) can be seen as a relationship between the "true" parameter and an initial value and the variance of σ_i is a measure of the analyst's "confidence" in the initial estimate. The two approaches coincide if the a priori estimates are obtained from sampling surveys.

To use traffic counts for estimating β, it is necessary to express the relationship linking these to the unknown parameters of a demand model. In general, to calibrate complete demand models, e.g. the sequence generation/distribution/ mode choice, it is necessary to have counts on the different modal networks. Let $\hat{f}_m$ be the vector of flows (measured) on the mode m network and $\hat{f}$ be the vector obtained by ordering sequentially all the vectors $\hat{f}_m$. The relationship between the traffic counts $\hat{f}_m$ and the "true" demand vector relative to mode m, $\boldsymbol{d}_m$, is basically analogous to (8.5.11) which now becomes:

$$\hat{\boldsymbol{f}}_m = \hat{\boldsymbol{M}}_m \boldsymbol{d}_m + \boldsymbol{\varepsilon}_m^{SIM} + \boldsymbol{\varepsilon}_m^{OBS} \tag{8.6.3}$$

The vector of O-D flows with mode m obtained through the demand model can be expressed as $\boldsymbol{d}_m(\boldsymbol{\beta})$, where, for simplicity sake, the vectors $\boldsymbol{SE}$ and $\boldsymbol{T}$ are understood[26]. Even if the "true" parameters vector $\boldsymbol{\beta}$ were known, the demand obtained from the model would diverge from the "true" demand by a vector of errors $\boldsymbol{\varepsilon}_m^{MOD}$:

$$\boldsymbol{d}_m = \boldsymbol{d}_m(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}_m^{MOD} \tag{8.6.4}$$

and by substituting (8.6.4) in (8.6.3) it results:

$$\hat{\boldsymbol{f}}_m = \hat{\boldsymbol{M}}_m \boldsymbol{d}_m(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}_m \tag{8.6.5}$$

where the vector $\boldsymbol{\varepsilon}_m$ is the sum of all the error components:

$$\boldsymbol{\varepsilon}_m = \boldsymbol{\varepsilon}_m^{SIM} + \boldsymbol{\varepsilon}_m^{OBS} + \hat{\boldsymbol{M}}_m \boldsymbol{\varepsilon}_m^{MOD}$$

and has zero mean if the vectors $\boldsymbol{\varepsilon}_m^{SIM}$, $\boldsymbol{\varepsilon}_m^{OBS}$, $\boldsymbol{\varepsilon}_m^{MOD}$ have zero mean.

The relationship (8.6.5) can be extended to the set of counting links belonging to different modal networks:

$$\hat{\boldsymbol{f}} = \hat{\boldsymbol{M}}\, \boldsymbol{d}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \tag{8.6.6}$$

where the vectors $\hat{\boldsymbol{f}}$, $\boldsymbol{d}(\boldsymbol{\beta})$ and $\boldsymbol{\varepsilon}$ are obtained by sequentially ordering the vectors of the different modes for which traffic counts are available. Similarly the assignment matrix $\hat{\boldsymbol{M}}$ is obtained by sequentially ordering modal assignment matrices.

The two information sources on $\boldsymbol{\beta}$, expressed respectively by equation (8.6.2) and (8.6.6), can be combined in different ways, obtaining different estimators of $\boldsymbol{\beta}$, in relation to the classic or Bayesian interpretation of the initial estimate $\hat{\boldsymbol{\beta}}$ and to the assumptions of the probability distribution of vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$. It is possible to specify Maximum Likelihood, Generalized Least Squares and Bayesian estimators of $\boldsymbol{\beta}$ analogous to those described in sections 8.5.1 and 8.5.2. Most estimators can be expressed in the general form:

$$\beta^* = \underset{b \in S_B}{argmin}\left[z_1(\boldsymbol{b}, \hat{\boldsymbol{\beta}}) + z_2(\hat{\boldsymbol{M}}\boldsymbol{d}(\boldsymbol{b}), \hat{\boldsymbol{f}})\right] \tag{8.6.7}$$

Note that the unknown parameters vector $\boldsymbol{b}$ has significantly less components than the O-D demand vector $\boldsymbol{x}$, (dozens of components instead of hundreds or thousand). Thus the problem (8.6.7) has a smaller dimensionality with respect to the ODCBE problem (8.5.12). Conversely, the optimization problem (8.6.7) is "more non-linear" than in the direct demand estimation case, since the non-linearity of demand models as function of unknown parameters is added to the non-linearity of functions $z_1(\cdot)$ and $z_2(\cdot)$.The feasibility set S_B may be coincident with the entire Euclidean space, as in the case of the Maximum Likelihood estimation dealt with in section 8.3, or constraints can be imposed on the "expected" signs of the coefficients (e.g. negative cost coefficients.)

Model (8.6.7) can also be specified when only aggregate information traffic counts or other sources is available. In this case the aggregate estimator results from the minimization of the "distance" $z_2(\cdot)$ between the observed traffic counts and the link flows obtained by assigning O-D flows generated by demand models. Unlike the ODCBE problem, it is possible to use only traffic counts since the number of independent counts is in general much larger than the number of unknown model parameters.

$$\beta^* = \underset{b \in S_B}{argmin}\; z_2(\hat{\boldsymbol{M}}\boldsymbol{d}(\boldsymbol{b}), \hat{f})$$

The Non-Linear Generalized Least Squares (*NLGLS*), is one of the most widely used specification of problem (8.6.7). This, in its simplified form, becomes:

$$\beta^* = \underset{b \in S_B}{arg\,min}\left[\sum_i \frac{(b_i - \hat{\beta}_i)^2}{Var[\sigma_i]} + \sum_l \frac{\left(\hat{f}_l - \sum_i \hat{m}_{li} d_i(\boldsymbol{b})\right)^2}{Var[\varepsilon_l]}\right] \tag{8.6.8}$$

Problem (8.6.8), can be solved by a gradient or a projected gradient algorithm similar to those described in Appendix A, according to whether constraints on the components of $\boldsymbol{b}$ have been specified or not. The k-th component of the gradient for objective function (8.6.8) can be expressed as:

$$GR_k = \frac{\partial}{\partial b_k}\left[\sum_i \frac{(b_i - \hat{\beta}_i)^2}{Var[\sigma_i]} + \sum_l \frac{(\hat{f}_l - \sum_i \hat{m}_{li} d_i(\boldsymbol{b}))^2}{Var[\varepsilon_l]}\right] = \tag{8.6.9}$$

$$= \frac{2(b_k - \hat{\beta}_k)}{Var[\sigma_k]} + 2\sum_l \frac{(\sum_i \hat{m}_{li} d_i(\boldsymbol{b}) - \hat{f})}{Var[\varepsilon_l]} \cdot \sum_i \hat{m}_{li} \frac{\partial d_i(\boldsymbol{b})}{\partial b_k}$$

The calculation of partial derivative of the demand function on the ith O-D pair with respect to the generic parameter β obviously depends on the specification adopted for the demand models being calibrated. Analytical calculation of these derivatives can be very cumbersome, or even impossible (e.g. for Probit models); in these cases recourse is made to numerical derivation methods.

The methods described have been applied to rather simple aggregate demand models, such as traditional four level models[27]. The results obtained are generally satisfactory. Fig. 8.6.1 shows an application of estimator (8.6.8) to the coefficients of a four-level demand model for the city of Reggio Calabria starting from two different vectors of initial parameters. In the literature there are no systematic comparisons of alternative specifications for $z_1(.)$ and $z_2(.)$.

From the statistical point of view model (8.6.7) can be considered as a two-stage mixed estimator (disaggregate/aggregate) of parameters β if it uses disaggregate information (choices of a users sample) to estimate $\hat{\beta}$, and aggregate information, traffic counts $\hat{f}$, for the correction of this initial estimate. It is also possible to formulate a "simultaneous" mixed estimator, such as a Maximum Likelihood estimator maximizing the probability of observing the choices $j(i)$ of a sample of users and the traffic counts $\hat{f}_l$. In this case, assuming that the observations are independent and that users' choices $j(i)$ are obtained with a simple random sample, the estimate β^{ML} can be obtained by combining the log-likelihood function (8.3.3) with one of the functions $z_2(\cdot)$ described in Fig. 8.5.4, expressing the log-likelihood of observing the counts as a function of the assignment matrix and of the parameters vector β:

$$\beta^{ML} = \underset{b \in S_b}{argmax}\left[\sum_i ln\, p^i[j(i)](\boldsymbol{b}) + z_2(\hat{\boldsymbol{M}}\boldsymbol{d}(\boldsymbol{b}), \hat{\boldsymbol{f}})\right]$$

The literature neither reports experiments with the simultaneous mixed estimator nor compares results with the sequential estimator.

	Model	Purpose	Attributes	$\hat{\beta}_1$	β^*	$\hat{\beta}_2$	β^*
β_1	Gener.	H-WPL	Workers	0.46	0.604	0.230	0.602
β_2	Gener.	H-SC	Students	0.86	0.902	1.015	0.900
β_3	Distrib.	H-WPL	Distance	1.02	0.346	1.103	0.347
β_4	Distrib.	H-WPL	Workplaces	0.70	0.570	1.008	0.550
β_5	Distrib.	H-SC	Distances	0.93	0.900	0.335	0.908
β_6	Distrib.	H-SC	School places	0.35	0.272	0.346	0.269
β_7	Mod. ch.	H-WPL	Walking Time	1.19	1.424	1.848	1.649
β_8	Mod. ch.	H-WPL	On-board Time	0.54	0.628	0.466	0.559
β_9	Mod. ch.	H-WPL	Cost car/bus	1.80	0.100	1.541	0.100
β_{10}	Mod. ch.	H-WPL	ASA Car	2.54	2.543	3.536	3.352
β_{11}	Mod. ch.	H-WPL	ASA Bus	2.29	2.330	2.116	3.179
β_{12}	Mod. ch.	H-SC	Walking time	2.18	2.207	3.436	2.737
β_{13}	Mod. ch.	H-SC	On-board Time	0.39	0.506	0.349	0.642
β_{14}	Mod. ch.	H-SC	Cost Bus	1.58	1.713	1.315	1.980
β_{15}	Mod. ch.	H-SC	ASA Bus	1.53	1.544	0.796	2.632

Demand model

$$d_{odm}(H-WPL) = \beta_1 Work_o \frac{exp[\beta_3 \ln \text{dist}_{od} + \beta_4 \ln \text{WPL}_d]}{\sum_{d'} exp[\beta_3 \ln \text{dist}_{od'} + \beta_4 \ln \text{WPL}_{d'}]} \cdot \frac{exp[V_{m/od}]}{\sum_{m'} exp[V_{m'/od}]}$$

$$d_{odm}(H-SC) = \beta_2 \text{Stud}_o \frac{exp[\beta_5 \ln \text{dist}_{od} + \beta_6 \ln \text{ScPL}_d]}{\sum_{d'} exp[\beta_5 \ln \text{dist}_{od'} + \beta_6 \ln \text{ScPL}_{d'}]} \cdot \frac{exp[V_{m/od}]}{\sum_{m'} exp[V_{m'/od}]}$$

Mode choice models

H-WPL $V_{walk} = \beta_7 T_w$

$V_{car} = \beta_8 T_c + \beta_9 Mc + \beta_{10} Car$

$V_{bus} = \beta_8 T_b + \beta_9 Mc + \beta_{11} Bus$

H-SC $V_{walk} = \beta_{12} T_w$

$V_{bus} = \beta_{13} T_b + \beta_{14} Mc + \beta_{15} Bus$

Number of counts

Road	: 30
Public transport	: 6
Pedestrians	: 26

Fig. 8.6.1 Example of demand model calibration with traffic counts.

A final consideration relates to path choice parameters. In all previous analyses, it has been assumed that the path choice model providing the elements $\hat{p}_{ki}$ of matrix $\hat{\boldsymbol{P}}$, and therefore the matrix $\hat{\boldsymbol{M}}$, was given i.e. that the parameters β^{PATH} in the systematic utility and in the random residuals distribution are known. These parameters were consequently not included in the vector β to be estimated. It is

possible to specify the estimation problem in order to improve an initial estimate $\hat{\beta}^{PATH}$ of these parameters by using traffic counts. In this case, the general expression of the model (8.6.7) becomes:

$$\beta^* = \underset{b \in S_b}{arg\,min}\left[z_1(\boldsymbol{b}, \hat{\beta}) + z_2(\hat{\boldsymbol{M}}(\boldsymbol{b})\,\boldsymbol{d}(\boldsymbol{b}), \hat{\boldsymbol{f}})\right] \tag{8.6.10}$$

where the vector β^{PATH} has been included in the general parameters vector β and in the variables vector $\boldsymbol{b}$. Comparing expressions (8.6.7) and (8.6.10), the latter is even more non-linear because the elements of the assignment matrix depend on unknown parameters.

A similar approach can be followed for the specification of joint estimators of O-D demand flows and path choice parameters. In this case the following formulation results:

$$\beta^*, \boldsymbol{d}^* = \underset{\substack{b \in S_b \\ x \in S_d}}{arg\,min}\left[z_1(\boldsymbol{x}, \hat{\boldsymbol{d}}) + z_2(\boldsymbol{b}, \hat{\beta}) + z_3(\hat{\boldsymbol{M}}(\boldsymbol{b})\boldsymbol{x}, \hat{\boldsymbol{f}})\right] \tag{8.6.11}$$

where the vector β coincides with β^{PATH}. Problem (8.6.11) simultaneously gives the estimates of path choice model parameters and demand flows minimizing the "distances" from their respective initial estimates and from the traffic counts observed.

Several other combined estimators of model parameters and/or demand flows can be specified along the lines described so far. It should be observed that the statistical properties and computational issues are at a very early research stage.

8.7. Estimation of intra-period dynamic demand flow using traffic counts

The O-D flows estimators discussed in section 8.5 were specified under the usual assumption of a within-day static system, i.e. that on average all relevant variables are constant within the reference period. In this section the statistical framework proposed for the static problem is generalized and extended to the dynamic O-D estimation case. This problem can be formally stated as that of combining time-varying traffic counts with other available information to estimate time-varying O-D flows. The problem is conceptually analogous to that discussed in section 8.5. The main difference is in the further complexity introduced by the within-day dynamic framework as discussed in Chapter 6. In this section some models developed for solving the Dynamic O-D Count Based Estimation (DODCBE) problem will be presented starting with the formal relationships between traffic counts and O-D flows. The DODCBE problem has been recently formulated in conjunction with the

inverse problem of Dynamic Traffic Assignment models (see Chapter 6), and it is by far less studied than its static counterpart.

Relationships between demand and counts.
Relationships between link flows and O-D flows will be expressed in the case of discrete time intervals that enable flows to be counted in practice.

Let the total study period J be divided into n_j intervals $j=1, \ldots, n_j$, of equal length T, so that $J=n_j T$. Let $d_{od}[j]$ represent the number of users moving between O-D pair o,d and leaving the origin during the interval j, $\boldsymbol{d}[j]$ the column vector obtained by arranging O-D flows. Let $\hat{d}_{od}[j]$ denote a priori information, or an initial estimate of the true demand $d_{od}[j]$,and $\hat{\boldsymbol{d}}[j]$ the corresponding vector.

For each interval j a link flow $\boldsymbol{f}_l[j]$ can be associated to each link l of the network[28], or more precisely to each section of a link, as the number of users crossing the section in that interval. In general, link counts over an interval are affected by measurement errors $\varepsilon_l^{OBS}[j]$. The measured flow $\hat{f}_l[j]$ is therefore only an estimate of the actual flow $f_l[j]$. In vector form:

$$\hat{\boldsymbol{f}}[j] = \boldsymbol{f}[j] + \varepsilon^{OBS}[j] \tag{8.7.1}$$

the link flow, $f_l[j]$ is comprised of O-D flows leaving during the same interval or in previous ones and reaching link l in interval j. This can be formally expressed by defining the quantity $m_{lj}^{od,t} \in [0,1]$ as the fraction of O-D flow $d_{od}[t]$ contributing to the flow on link l in interval j, resulting in:

$$f_l[j] = \sum_{t=1}^{j} \sum_{od} m_{lj}^{od,t} d_{od}[t] \tag{8.7.2}$$

Equation (8.7.2) can be expressed in matrix form by introducing the $(n_l \times n_{od})$ assignment fraction matrices $\boldsymbol{M}[t,j]$, analogous to the within-day static counterpart defined in equation (8.6.3):

$$\boldsymbol{f}[j] = \sum_{t=1}^{j} \boldsymbol{M}[t,j]\boldsymbol{d}[t]$$

This equation assumes that demands flows and counts before the first interval are negligible. This assumption introduces a positive bias in O-D estimates for the first interval; it can be easily relaxed if an estimate of O-D demand leaving before the study period is available.

Let $h_k[j]$ be the path flow, i.e. the average number of travelers per time unit following path k between O-D pair o,d and leaving during period j. Path flows can

also be expressed as the product of the O-D demand $d_{od}[t]$ and the probability (average fraction) $p[k/t]$ of choosing path k given the departing interval t:

$$h_k[t] = d_{od}[t] \cdot p[k/t] \tag{8.7.3}$$

In order to express assignment fractions $m_{lj}^{od,t}$ in terms of path choice probabilities, the formal dependence of link flows on path flows must be introduced:

$$f_l[j] = \sum_{od} \sum_{k \in K_{od}} \sum_{t=1}^{j} b_{lj}^{kt} h_k[t] \tag{8.7.4}$$

where the summation is extended to all paths belonging to the set K_{od} of paths connecting O-D pair o,d.

In the above expression b_{lj}^{kt} is the crossing fraction, i.e. the fraction of path flow $h_k[t]$ crossing a section of link l at interval j; the above fractions depend on how link flows are defined, when each path flow reaches link l, and how it moves on it.

By combining equation (8.7.3) and (8.7.4), and comparing with equation (8.7.2), it results:

$$m_{lj}^{od,t} = \sum_{k \in K_{od}} b_{lj}^{kt} p[k/t] \tag{8.7.5}$$

Equation 8.7.4 can be expressed in matrix form:

$$\boldsymbol{f}[j] = \sum_{t=1}^{j} \boldsymbol{B}[t,j] \boldsymbol{h}[t]$$

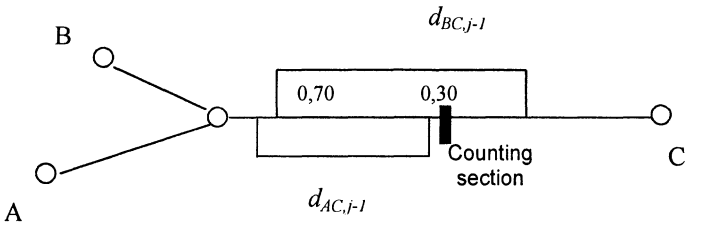
where $\boldsymbol{B}[t,j]$ is the crossing fraction matrix $\boldsymbol{B}[t,j] = \left\{ b_{lj}^{kt} \right\}$.

In practice, only estimates $\hat{p}[k/t]$ and $\hat{b}_{lj}^{kt}$ of the true values $p[k/t]$ and b_{lj}^{kt} can be obtained through path choice and Dynamic Network Loading (DNL) models, see Chapter 6. Estimates of assignment fractions can thus be formally expressed as:

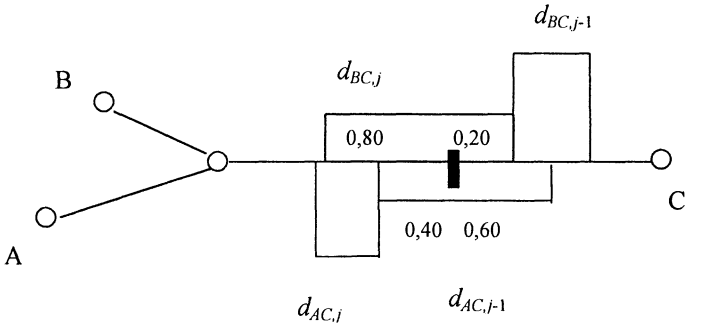
$$\hat{m}_{lj}^{od,t} = \sum_{k \in I_{od}} \hat{b}_{lj}^{kt} \hat{p}[k/t] \tag{8.7.6}$$

Crossing fractions are included in the interval [0,1] if path flows $h_k[j]$ are modeled as space-continuous packets. In the most frequent models of space-discrete packets, crossing fractions are either 0 or 1 depending on whether packet $[k,j]$ crosses the counting sections on link l during interval t.

Fig. 8.7.1 shows an elementary example of the relationship between within-day dynamic traffic counts and O-D demand flows, both in case of space-continuous path flows and space-discrete path flows.



Beginning of interval j – flows spatial positions



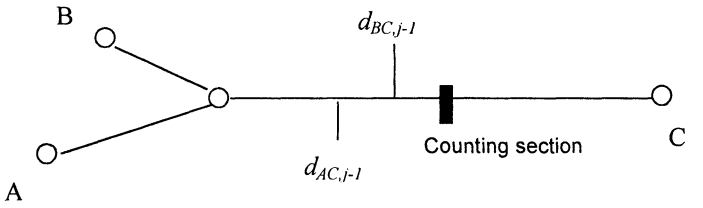
End of interval j – flows spatial positions

$$m_{lj}^{AC,j-1} = 0.60 \qquad m_{lj}^{BC,j-1} = 0.70$$

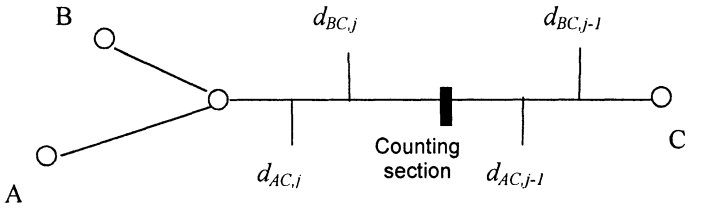
$$m_{lj}^{AC,j} = 0 \qquad m_{lj}^{BC,j} = 0.20$$

$$f_l[j] = 0.60\, d_{\text{AC, j - 1}} + 0.70 d_{\text{BC, j - 1}} + 0.20 d_{\text{BC, j}}$$

Fig. 8.7.1a Relationship between within-day dynamic traffic counts and O-D flows – continuous path flows representation.



Beginning of interval j – packets spatial positions



End of interval j – packets spatial positions

$$m_{lj}^{AC,j-1} = 1 \qquad m_{lj}^{BC,j-1} = 1$$

$$m_{lj}^{AC,j} = 0 \qquad m_{lj}^{BC,j} = 0$$

$$f_l[j] = d_{AC,j-1} + d_{BC,j-1}$$

Fig. 8.7.1b Relationship between within-day dynamic traffic counts and O-D flows – discrete path flows representation.

The estimated values of crossing fraction $\hat{b}_{lj}^{kt}$, path choice probabilities $\hat{p}\,[k/t]$ and, consequently, assignment fractions $\hat{m}_{lj}^{od,t}$ are expected to be different from the true ones. This implies, as already seen in the static context, that even if the true demand vector $\boldsymbol{d}[t]$ were known and assigned to the network substituting $\hat{m}$ instead of m in equation (8.7.2), the resulting link flows would differ from the actual ones by a random error term ε^{SIM} (modeling error):

$$f_l[j] = \sum_{t=1}^{j} \sum_{od} \hat{m}_{lj}^{od,t} d_{od}[t] + \varepsilon^{SIM}[j] \tag{8.7.7}$$

or in matrix form:

$$f[j] = \sum_{t=1}^{j} \hat{\boldsymbol{M}}[t, j]\boldsymbol{d}[t] + \varepsilon^{SIM} \tag{8.7.8}$$

Equation (8.7.1) and (8.7.8) can be combined into:

$$\hat{f}[j] = \sum_{t=1}^{j} \hat{\boldsymbol{M}}[t, j]\boldsymbol{d}[t] + \varepsilon \tag{8.7.9}$$

where the random vector ε is the sum of the two (independent) vectors ε^{OBS} and ε^{SIM}.

Dynamic estimation of O-D demand flows.
In section 8.5, it has been demonstrated that most O-D demand static estimators can be obtained by solving a constrained optimization problem of the form:

$$\boldsymbol{d}^* = \underset{x \in S}{arg\,min}\left[z_1\left(\boldsymbol{x}, \hat{\boldsymbol{d}}\right) + z_2\left(v(\boldsymbol{x}), \hat{\boldsymbol{f}}\right)\right] \tag{8.7.10}$$

In this section the estimators previously proposed for the static context will be extended to the two dynamic estimation cases.

The problem here is to estimate O-D demand flows for each interval, $\boldsymbol{d}[t]$, by using counts $\hat{\boldsymbol{f}}[j]$. Two alternative approaches are possible. The first, referred to as *simultaneous*, looks for an estimator giving, in a single step, the whole O-D demand pattern ($d[1]$, ..., $d[n]$) by using simultaneously counts over all intervals. The second approach, referred to as *sequential*, produces at each step the O-D demand vector for one period, using counts relative to that period and to the previous one and, possibly, O-D estimates relative to previous periods. In the following subsections the two estimators will be described separately.

8.7.1. Simultaneous Estimators

Static estimators can be extended in a straightforward manner to the simultaneous estimation framework. In this case, however, the single unknown demand vector has to be replaced by the n_j vectors ($\boldsymbol{x}[1]$, ..., $\boldsymbol{x}[j]$, ..., $\boldsymbol{x}[n_j]$). Likewise, the counted flow

vector is replaced by $\left(\hat{f}[1],...\hat{f}[j],...\hat{f}[n_j]\right)$. The general form of the estimators then becomes:

$$d^*[1]...d^*[n_j] = \underset{x[1]\geq 0,...,x[n_j]\geq 0}{argmin}\left[z_1\left(x[1],...,x[n_j];\hat{d}[1],...,\hat{d}[n_j]\right)+ z_2\left(x[1],...,x[n_j];\hat{f}[1],...,\hat{f}[n_j]\right)\right] \tag{8.7.11}$$

All specifications of objective functions $z_1(\cdot)$ and $z_2(\cdot)$ reported in sections 8.5.1 and 8.5.2 can be extended and substituted in equation (8.7.11) thus obtaining ML, GLS, or Bayesian estimators depending on the distribution assumptions made on the random residuals ε_j. For example, in the case of GLS estimator, they become:

$$z_1 = \sum_{j=1}^{n}\left(x[j]-\hat{d}[j]\right)^T Z^{-1}[j]\left(x[j]-\hat{d}[j]\right)$$

$$z_2 = \sum_{j=1}^{n}\left(\sum_{t=1}^{j}\hat{M}[t,j]x[t]-\hat{f}[j]\right)^T W^{-1}\left(\sum_{t=1}^{j}\hat{M}[t,j]x[t]-\hat{f}[j]\right)$$

8.7.2. Sequential Estimators

In this context, an O-D demand vector is estimated for a single interval j at each time interval. There are two advantages to this approach. The first is the reduction of computational complexity by decomposing a large optimization problem into a number of smaller and more manageable ones; the second is that estimates obtained for an interval can be used as initial estimates in subsequent estimations.

The main idea in this approach is to express counts of a period j as a linear (stochastic) function of the unknown demand of the same period only. This is achieved by equating the demand relative to previous periods to the already computed estimates $d^*[t]$:

$$\hat{f}[j] = \sum_{t=1}^{j-1}\hat{\mathbf{M}}[t,j]\boldsymbol{d}^*[t]+\hat{\mathbf{M}}[j,j]x[j]+\varepsilon[j] \tag{8.7.12}$$

The general formulation of the static estimation problem can be adapted to this context, leading to:

$$d^*[j] = \underset{x[j]\geq 0}{\arg\min}\left[z_1\left(x[j],\hat{d}[j]\right)+ z_2\left(x[j]/d^*[1],...,d^*[j-1];\hat{f}[j]\right)\right] \tag{8.7.13}$$

where $\hat{f}[j]$ is given by 8.7.12..

In the case of GLS estimator, the objective functions z_1 and z_2 become:

$$z_1 = \left(x[j] - \hat{d}[j]\right)^T Z^{-1}[j]\left(x[j] - \hat{d}[j]\right)$$

$$z_2 = \left(\sum_{t=1}^{j-1} \hat{M}[t,j]d^*[t] + \hat{M}[j,j]x[j] - \hat{f}[j]\right)^T W^{-1}\left(\sum_{t=1}^{j-1} \hat{M}[t,j]d^*[t] + \hat{M}[j,j]x[j] - \hat{f}[j]\right)$$

8.8. Applications of demand estimation methods

The methods described in this chapter can be used to estimate present demand flows or demand flows corresponding to hypothetical scenarios for the transportation system and/or for the activity system. These estimates can in turn be used to simulate link flows and performances with an assignment model and/or to analyze the structure of the transportation demand in a given area. Obviously, different techniques, or combinations of techniques, can be used for different applications and for different components of the demand. In the following the main areas of application and the relative demand estimation methodologies will be described, summarizing the results of the previous sections (see Fig. 8.8.1)

Area of application	Estimation method	Input data	Complementary tecniques
Estimation of present demand	Direct estimation	Sampling surveys	Estimation of O-D matrices with traffic counts
	Model estimation	• Models parameters • Attributes of the activity systems SE^P • Attributes of the transport system T^P	
Estimation of demand variations (forecast)	Model estimation	• Models parameters • Attributes of the activity systems (Scenarios) SE^F • Attributes of the transport system (Projects) T^F	Pivoting on the present demand

Fig. 8.8.1 Application of demand estimation methods.

8.8.1. Estimation of present demand

The estimation of average demand flows in the reference period can be performed using sampling surveys and direct estimation methods, or by applying a system of demand models to the present configuration of the system.

In the former case, the sampling methods described in section 8.2 are used. From the practical point of view, it should be noted that different types of sampling surveys are often used for the estimation of different components of the demand. In particular, on-board or en-route surveys are often used to estimate exchange and crossing flows while household surveys are used to estimate internal demand flows.

Demand models can be used as estimators of present demand by applying them with present values of the attributes of the activities system, $\boldsymbol{SE}^P$, and of the

transportation supply system, $\boldsymbol{T}^P$. Model estimation of present demand can be formally expressed as:

$$\hat{\boldsymbol{d}}^P_{MOD} = \boldsymbol{d}(\boldsymbol{SE}^P, \boldsymbol{T}^P; \hat{\boldsymbol{\beta}}) \tag{8.8.1}$$

where $\hat{\boldsymbol{\beta}}$ indicates the estimate of the parameters vector. Expression (8.8.1) can be applied to estimate demand flows with different levels of aggregation, e.g. by origin, destination and mode.

Model-based estimation of present demand deserves a few comments.

- The rationale of the method is that, for a given sample size, estimates of the parameters $\hat{\boldsymbol{\beta}}$ are significantly more precise than direct sampling estimates of $\hat{d}$. The underlying assumption of the method is that deviations between true demand flows and model-based estimates are less dispersed than the deviations between direct estimates and the true demand flows. This assumption has received some, though limited, empirical validation.
- The application of demand models requires the aggregation of the results. The different aggregation techniques described in section 3.7 can be used to obtain estimates $\hat{\boldsymbol{d}}$ of the trips flows between the different origin-destinations pairs. Aggregation by categories (aggregate models) and sample enumeration (disaggregate models) are the most common options.
- The models used for present demand estimation might be different and less sophisticated than those used to predict demand variations. In the former case, in fact, the model should be seen as "estimator" with the exclusive function of reproducing, descriptively, the observed phenomenon. On the other hand it is reasonable to assume that a model with interpretative capabilities should be a better "predictor". Again, models of various levels of complexity can be used to estimate different components of present demand. In particular, exchange demand can be estimated with simpler models requiring less information than those used to estimate demand flows within the study area.
- Model specification, calibration and validation can be conducted using the disaggregate methodologies described in sections 8.3 and 8.4, possibly integrated with the mixed aggregate/disaggregate estimation method using traffic counts described in section 8.6.

The two methods (direct estimation and model-based estimation) are generally used to estimate different components of present demand. For example, it is quite common to use direct estimation for exchange and crossing demand (for which it is at the same time easier to conduct direct cordon surveys and more complicated to formulate demand models) and model-based estimation for internal demand. Finally, present demand can be estimated by combining direct estimation and/or

model-based estimation with aggregate information on traffic counts using the methods described in section 8.5.

8.8.2. Estimation of demand variations (forecasting)

The classic use of demand models is to simulate demand variations following modifications of the activity system and/or of the transportation supply system. There is obviously a close interdependence between the characteristics of demand models and the project under study, since the model must be "elastic" with respect to variables describing the changes whose effects are to be evaluated. For example, for the circulation plan of an urban road network, it is sometimes assumed that the transportation demand, with all its characteristics, remains unchanged except for the users' path choices. This implies that the present O-D demand matrix for the "car" mode can be used to simulate the consequences of alternative projects and that the only demand model necessary for this purpose is the path choice used for rigid demand assignment. On the other hand, if the same plan is included in a wider project aimed at modifying the modal split of present demand, e.g. by introducing park pricing, it will be necessary to use modal choice and path choice models, which can be applied to present O-D matrices.

In general, in the case of *short-term projects,* it is assumed that the socio-economic variables of the activity system remain unvaried while the transportation performance variables are modified by the project. These variations may impact travel choices on several dimensions (path, mode, destination, frequency). In this case, the application of the demand models can be formally expressed as:

$$\hat{\boldsymbol{d}}^{F}_{MOD} = \boldsymbol{d}(\boldsymbol{SE}^{P}, \boldsymbol{T}^{F}; \hat{\boldsymbol{\beta}}) \tag{8.8.2}$$

where $\hat{\boldsymbol{d}}^{F}_{MOD}$ indicates the vector of the model-based estimates of "future" demand flows and $\boldsymbol{T}^{F}$ indicates the vector of level-of-service attributes corresponding to the project.

Medium-long term projects, usually require the simulation of their effects over a sufficiently long period. In this case it is it is necessary to forecast the evolution of these variables. In general it is very difficult to forecast the evolution of the main variables of the activity system such as resident population, incomes levels, economic production organization and life styles of families, location of manufacturing and services activities. These are significant factors that are difficult to forecast reliably in the medium-long term. Even if some variables of the activity system can be considered endogenous in the models system, particularly in the case of transport-land use interaction models, the evolution of several other exogenous variables still has to be forecasted. In practice, for long-term applications, different scenarios[29] for the evolution of the variables $\boldsymbol{SE}^{F}$ are used. Demand models are applied to each scenario and variation ranges of the key variables can be used for the design and the evolution of alternative projects as will be seen in Chapter 10. The

estimation of demand flows over long periods can therefore be formally expressed as:

$$\hat{\boldsymbol{d}}^F_{MOD} = \boldsymbol{d}(\boldsymbol{SE}^F, \boldsymbol{T}^F; \hat{\boldsymbol{\beta}}) \tag{8.8.3}$$

The comments on model calibration and aggregation techniques can be extended to both the applications (8.8.2) and (8.8.3).

Recently, forecasting techniques alternative to those expressed by (8.8.2) and (8.8.3) are sometimes used. These techniques are based on the "pivoting" method in which models are used to estimate the variations with respect to present demand, rather than directly future demand. This approach assumes that it is possible to obtain estimates, $\hat{\boldsymbol{d}}^P$, of the present demand "better" than those obtained by using only demand models. This may be the case if other information sources on present demand are available, so that direct or model-based estimates of present demand are improved with that information (e.g. traffic counts). In this case, modeling errors can be reduced by using the models as simulators of demand variations and, therefore, obtaining "future" demand estimates as:

$$\hat{d}^F_{od} = \hat{d}^P_{od} \cdot \frac{d_{od}(\boldsymbol{SE}^F, \boldsymbol{T}^F; \hat{\boldsymbol{\beta}})}{d_{od}(\boldsymbol{SE}^F, \boldsymbol{T}^P; \hat{\boldsymbol{\beta}})} \tag{8.8.4}$$

The general form (8.8.4) must be specialized for the demand dimensions to which it is applied. For example, by applying the method to modal O-D matrices and leaving to the network assignment the definition of future path choice probability and the resulting flows (see Fig. 8.8.2), the expression (8.8.4) becomes:

$$d^F_{od}[shm] = \hat{d}^P_{od}[shm] \cdot \frac{d_{od}[shm](\boldsymbol{SE}^F, \boldsymbol{T}^F; \hat{\boldsymbol{\beta}})}{d_{od}[shm](\boldsymbol{SE}^P, \boldsymbol{T}^P; \hat{\boldsymbol{\beta}})} \tag{8.8.5}$$

The application of the pivoting method in the form of (8.8.4) requires a double application of the model to present ($\boldsymbol{SE}^P$, $\boldsymbol{T}^P$) and future ($\boldsymbol{SE}^F$, $\boldsymbol{T}^F$) scenarios. Furthermore, the method must be adapted for practical applications; for example, equation (8.8.5) would not allow the estimation of demand associated with the introduction of a new mode of transport for which no present demand exists. These distortions can be corrected in various ways, for example by applying the pivoting method partially to simulate variations of the present demand only on some dimensions and then applying directly the models to the other dimensions.

OD Pair	$\hat{d}^{P}_{od\,Car}$	$\hat{d}^{P}_{od\,Train}$	$d^{P(MOD)}_{od\,Car}$	$d^{P(MOD)}_{od\,Train}$	$d^{F(MOD)}_{od\,Car}$	$d^{F(MOD)}_{od\,Train}$	$\hat{d}^{F}_{od\,Car}$	$\hat{d}^{F}_{od\,Train}$
1,2	100	30	92	31	85	40	92.4	38.7
1,3	30	15	26	11	22	25	25.4	34.1
1,4	70	25	73	22	60	31	57.5	35.2
2,1	120	46	116	47	103	53	106.6	51.9
2,3	50	22	47	19	49	29	52.1	33.6
2,4	60	18	55	20	53	31	57.8	27.9
3,1	85	32	88	27	76	39	73.4	46.2
3,2	70	27	71	30	68	46	67.0	41.4
3,4	23	5	20	6	18	11	20.7	9.2
4,1	58	24	56	22	52	30	53.9	32.7
4,2	65	26	66	24	60	35	59.1	37.9
4,3	90	32	87	33	70	48	72.4	46.5

Fig. 8.8.2 Application of the pivoting method.

Reference Notes

Direct demand estimation is based on the application of sampling surveys and estimators. A description of "classical" travel demand surveys can be found in manuals such as the one from RRL (1965) and DOT, EPA (1996). For statistical sampling theory, refer to the texts by Cochran (1963) and Yates (1981). Applications to transportation demand estimation are covered in several articles, such as those of Smith (1979) and Brog and Ampt (1982) as well as in the volume by Ortuzar and Willumsen (1994)

The literature on specification, calibration and validation of demand models is quite substantial. The books by Domenicich and McFadden (1975) and Ortuzar and Willumsen (1994), as well as the articles by Horowitz (1981), (1982), Manski and McFadden (1981) address various statistical aspects of the calibration of disaggregate models. A review of the field at the date is contained in Gunn and Bates (1982). The work of Manski and Lerman (1977) studies model calibration based on non-random samples. A detailed and systematic discussion of the subjects in section 8.3 is contained in the volume by Ben Akiva and Lerman (1985) and the reader is referred to the latter's comprehensive bibliography.

Stated Preferences survey techniques have been the object of growing interest over the last 10-15 years and are an area in continuous evolution both from the theoretical and from the application point of view. A discussion of the theoretical aspects of SP techniques can be found in the works of Hensher et al. (1988), Louviere (1988), and Ortuzar (1992), while practical aspects are covered in Pearmin et al. (1991). The statistical bases of factorial analysis are described in greater detail in the texts on experimental design such as Box and Hunter (1978). The combined

calibration of demand models on the basis of SP-RP surveys is dealt with in Ben Akiva and Morikawa (1990) and Bradley and Daly (1992). An application to mode choice modeling is described in Biggiero and Postorino (1994), the example in fig. 8.4.5 is reported there.

The estimation of demand flows using traffic counts is a subject intensely researched over the last two decades. An updated state of the art and literature review can be found in Cascetta and Improta (1999). The general statistical bases are addressed in Cascetta and Nguyen (1986). For estimation of O-D demand flows using traffic counts, there are several papers on particular estimators or specific applications. The papers by Van Zuylen and L.G. Willumsen (1980) on Maximum Entropy estimator, Maher (1983) on Bayesian estimators, Cascetta (1984) proposing the GLS estimator, Bell (1991) on applications of the GLS method, Di Gangi (1988) on numerical comparison of the statistical "performances" of different estimators can be quoted.

The problem of estimating O-D flows using traffic counts in congested networks is relatively more recent; it has been studied by a number of authors, typically as a bi-level programming problem for DUE assignment, see Chen and Florian (1995), Yang (1995). The fixed–point formulation and the MSA algorithm described in section 8.5 with some variants, are described in Cascetta and Postorino (2000).

Estimation of model parameters using traffic counts and other sources is a well established heuristic practice, but has received relatively limited attention from the theoretical point of view. Among the first papers proposing methods for aggregate estimation of coefficients using traffic counts, can be quoted those by Cascetta (1986) proposing GLS estimators and by Willumsen and Tamin (1989) describing an estimator for gravity type models. The paper by Cascetta and Russo (1997) describes the general statistical framework discussed in section 8.6. The combined estimation (both aggregate and disaggregate) of model parameters and O-D flows using traffic count is original.

In the literature, many authors have proposed different methods to estimate time-varying O-D flows using traffic counts. Among others, Cremer and Keller (1987) propose sequential estimators in the case of a simple network using traffic counts only. Cascetta, Inaudi and Marquis (1993) propose dynamic estimators obtained by optimizing a two-term objective function as described in section 8.7. Nguyen, Morello and Pallottino (1986) proposed different simultaneous estimators on a general transit network. Okutani and Stephanades (1984) use a standard Kalman Filtering approach in order to obtain sequential estimators, while Ashok and Ben-Akiva (1993) prove that the same approach can be used to obtain GLS simultaneous estimators on deviations.

Notes

(1) For simplicity of notation in the following, relevant characteristics of demand flows, e.g. *s h m* and the users category, are understood.

(2) The coefficient (1-α), known as the finite population correction coefficient, accounts for the fact that the population has a finite number of members; therefore, if a census, α=1, were conducted, the estimate would be the "true" value with a zero variance. The weight of the correction coefficient, however, is negligible for the sampling rates used in practice for direct demand estimation.

(3) A different method, which can be referred to as aggregate calibration, uses aggregate and indirect information on users' travel behavior, usually traffic counts, to specify and calibrate demand models. There are also mixed methods which use disaggregate and aggregate information simultaneously. Aggregate and mixed estimators of demand models parameters will be covered in section 8.6

(4) Note that terms of the summations in equations (8.3.7) and (8.3.8) are the difference between the systematic utility for the chosen alternative, $V_{j(i)}(\boldsymbol{X}, \boldsymbol{\beta})$, and the satisfaction associated with all the available alternatives:

$$ln\, L(\boldsymbol{\beta}')=\Sigma_{i=1...n}[V_{j(i)}(\boldsymbol{X}^i, \boldsymbol{\beta}) - s(\boldsymbol{X}, \boldsymbol{\beta})]$$

and this difference, as can be seen from section 3.5, is always less than zero.

(5) Note that the estimates of a given coefficient β_k obtained with different specifications of the random utility model (Multinomial Logit, Hierarchical Logit, Probit) are usually different since they contain different scale coefficients.

(6) For simplicity of notation in what follows, no distinction will be made between the vector $\boldsymbol{\beta}$ of the coefficients in the utility function and the vector $\boldsymbol{\theta}$ of the structural coefficients, or more properly between the vectors $\boldsymbol{\beta}'$ and $\boldsymbol{\delta}'$ of the identifiable parameters. The vector $\boldsymbol{\beta}$ is to be understood as the set of all the coefficients to be estimated.

(7) In theory, the model's goodness of fit should be tested on a sample of observations different from the sample used for the calibration (hold-out sample). In practice, this procedure is not always followed to make the best use of all the available information, given the limited size of many available samples.

(8) This type of assumption is known as "non-nested".

(9) Choice alternatives in any scenario depend on the functional form of the model to be calibrated. With Multinomial Logit models, due to the IIA property, estimates of the systematic utility coefficients do not depend on the number of alternatives proposed, so that the scenario might include any subset of the alternatives included in the model. In the case of models for which the IIA is not valid, e.g. Hierarchical Logit and Probit, scenarios must be designed to account for the structure of the model explicitly. For example, alternatives belonging to different groups, as well as multiple alternatives for the same group, must be included in some scenarios for Hierarchical Logit models.

(10) These techniques are derived from multivariate statistical analysis and, in particular, from experimental design techniques. They are designed to allow the analysis of direct and indirect effects of relevant variables by means of linear models and therefore do not correspond exactly to the case of demand models, which are typically non-linear with respect to explicative variables (attributes).

(11) In practice, it would be more correct to assume the existence of a correlation between the observations relative to the choices of each individual. In this case, however, the expression of the log-likelihood function would be considerably more complicated; for this reason the correlation effects are usually ignored in applications.

(12) This is, at the same time, the most frequent and most complex case. Other aggregate information sources can be easily represented as particular cases of link counts by properly specifying the "assignment equation" (8.5.2). Total generated and/or attracted flows, average trip length, distribution of trip lengths, and total flows crossing internal cordons are examples of other aggregate information on O-D flows which can be seen as special cases. In the following it will also be assumed that flow-counting locations are given, i.e. the links are given as input to the problem of O-D demand estimation. Although this is sometimes the case, counts location should be designed with respect to their information content. The problem of optimal counting locations can be formulated as a network design problem similar to those described in Chapter 9.

(13) For simplicity of notation in this section, the generic element of the demand vector will be denoted as d_i, i. e., using a single index for the O-D pair as in Chapter 5, instead of the double index d_{od} used previously.

(14) Note that values of p_{ki}, and therefore m_{li}, are the "true" values, i.e. the true fractions of users who use a given path or a given link in the reference period.

(15) The calculation of the assignment matrix $\hat{\boldsymbol{M}}$ in the case of congested networks for which the link costs are not known will be covered in section 8.5.4 relative to computational aspects.

(16) Note that expressions (8.5.6) and (8.5.7) correspond to the relationships defining road networks assignment models and public transport assignment models respectively. The main difference lies in the explication of the approximations connected to the assignment model in the vector $\boldsymbol{\varepsilon}^{SIM}$. Also, to highlight this difference, in this section, $\boldsymbol{v}$ will indicate the vector of the link flows resulting from the assignment of the "true" demand vector $\boldsymbol{d}$ while in Chapter 5 this vector is indicated by $\boldsymbol{f}$.

(17) It should be remembered that the components of the "true" vectors $\boldsymbol{d}$ and $\boldsymbol{f}$ are the flows between each O-D pair and each link average over different observation periods.

(18) This can be seen as a confirmation of the essentially interpretative nature of the difference between the objective and subjective approaches in probability theory and statistical inference.

(19) The equation (8.5.16) is obtained from (8.5.15) by using the Stirling approximation: $ln(x!) \cong x\ ln\ x\text{-}x$.

(20) Bayesian estimators coincide with "classic" estimators under the assumption that subjective estimates are also obtained from sampling surveys. This indicates that "classic" estimators can be obtained as special cases in the context of Bayesian statistics.

(21) In theory, it is possible to derive different Bayesian estimators corresponding to the parameters of the a posteriori probability function. The estimator given by (8.5.27) corresponds to the mode of the a posteriori probability function (8.5.26). Another estimator could be obtained as the expected a posteriori vector:

$$\boldsymbol{d}^{B} = E[x/\hat{\boldsymbol{f}}, \hat{\boldsymbol{d}}\,] = \int_{x \in S} x h(x \quad \hat{f}, \hat{d}) dx$$

In practice, however, the calculation of this estimator would be very complex since it is not usually possible to solve analytically the multiple integral defining it.

(22) The problem (8.5.34) can be easily applied to maximization problems by changing the sign of the objective function.

(23) In the literature, the fixed-point formulation has been proposed mainly for SUE assignment, where the equilibrium assignment map is defined uniquely, and the bi-level formulation has been proposed with reference to DUE assignment.

(24) The methods described in this section, although presented with reference to traffic counts, can easily be extended to mixed (aggregate/disaggregate) or purely aggregate calibration, using other aggregate data. For example, the parameters can be estimated on the basis of estimates $\hat{d}_{odl}$ of demand flows derived from different sources (data from transport companies or sampling estimates). In this case the assignment matrix $\hat{\boldsymbol{M}}$ relating the aggregate counts to the demand vector is the identity matrix. Other aggregate data can complement or substitute traffic counts.

(25) The vector $\boldsymbol{\beta}$ denotes all the identifiable parameters of the specific demand models system, including those relative to the random residuals probability density function. It will include, for example, the coefficients β'_i of a Multinomial Logit model and the coefficients β'_i and δ_j of a Hierarchical Logit model.

(26) Note that the O-D demand on a given mode m usually depends on the level of service attributes for all the competing modes. For this reason the vector $\boldsymbol{d}_m$ has been expressed as a function of the vector $\boldsymbol{T}$ including the attributes of all transport modes. Furthermore, since it is assumed that the assignment matrix $\hat{\boldsymbol{M}}$ is known, the vector of the unknown parameters $\boldsymbol{\beta}$ does not include those relative to path choice. This assumption will be relaxed in what follows.

(27) In the case of disaggregate demand models, and sample enumeration aggregation techniques, all the previous expressions still hold. However, each calculation of the demand flows vector requires the application of the entire aggregation procedure, which might be quite burdensome.

(28) In Chapter 6 it was stated that in a within-day dynamic system user's flows may vary for different cross-sections of the same link. The generic flow at section s of link l was denoted by $f_{l,s}[0]$. In the following it will be assumed that only one counting section is associated to a counted link, the link flow is thus relative to that counting section. Furthermore, to be consistent with the notation of the static case, unlike in Chapter 6, the generic link will be denoted by l.

(29) A scenario can be defined as a set of internally consistent assumptions on the exogenous variables of a models system. In some applications scenarios are obtained with other macro-economic models requiring less input variables (e.g. population and economic growth rates), scenario models generate consistent sets of disaggregate input variables for transportation demand models.

9 TRANSPORTATION SUPPLY DESIGN MODELS

9.1. Introduction

This chapter outlines a wide range of methods and mathematical models which may assist the transportation systems engineer in designing projects or interventions. It should be stated at the outset that supply design models[1] are not meant to "automate" the complex task of design, especially when the proposed actions can alter significantly the performances of the transportation system. In this case, as we have seen, the project may have structural effects ranging from changes in land use to modifications in the level and structure of travel demand. On the other hand, the elements of the transportation supply to be designed may assume a very large number of possible configurations; circulation directions in an urban road network or the lines and frequencies of a transit system are two cases in point. In these cases it is practically impossible to explore and compare all the feasible configurations to identify the optimum with respect to a given set of objectives and constraints.

From the modeling perspective, supply design models belong to a different class than the models described so far and, in some respect, can be considered as extensions or generalizations of these models. The mathematical models described in the previous chapters, in fact, aim at simulating the relevant aspects of a transportation system under the assumption that supply (facilities, services and prices) and activity systems are exogenously given. These models can be used as "design tools" by simulating the main effects of exogenously specified projects, verifying their technical compatibility and evaluating their "convenience" as will be seen in Chapter 10. This approach is known as "what if". On the other hand, supply design models provide "what to" indications, i.e., how to alter supply in order to optimize given objectives while satisfying given constraints (see Fig. 9.1.1). Clearly, in order to identify solutions for the design problem, it is necessary to evaluate the system responses (demand, flows and performances) to the possible actions; therefore the simulation model is a component of the design model. The cost of this generalization is not only the simplification of the real design problem, but also the simplification of the simulation models, now sub-models of a wider model.

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

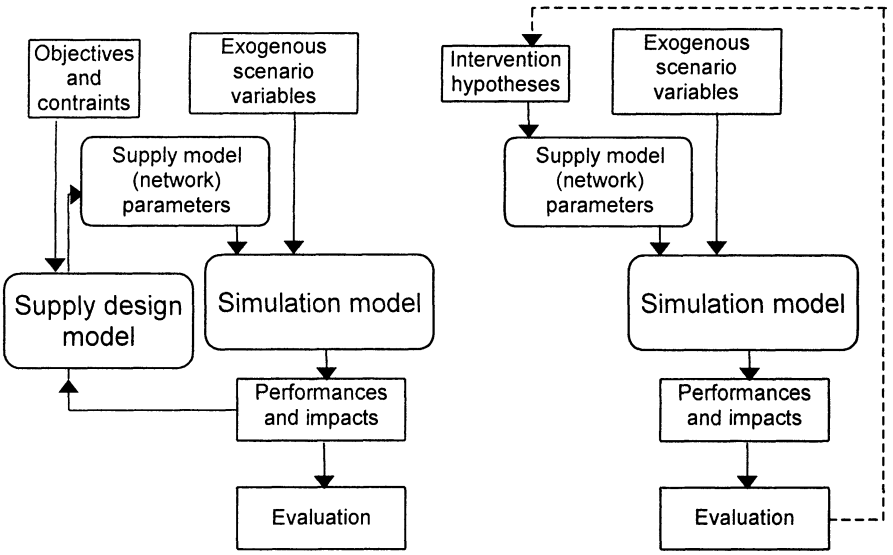


Fig. 9.1.1 Two approaches to the design of transportation supply.

An interesting interpretation of the differences between simulation and design models can be given in terms of game theory. The design problem can be seen as a "Stackelberg game". One of the two players (or groups of players), called the *leader*, knows in advance the reactions of the other player (or group of players), called the *follower*, to his/her *actions*. In this case, the leader is the designer (or manager) of the supply system and the followers are the users of the transportation system. The designer is able to anticipate the reactions of the users and exploits this information to achieve his/her objectives[2]. In this context, the simulation models represent the tools to predict users' reactions, while supply design models provide the leader with the "winning" strategy. On the other hand, within the context of the game theory, the simulation model can be interpreted as a description of a "Nash game", in which the generic player (say users) ignores the possible reactions of the other players.

Supply design models typically simplify the actual design problem, accounting for only some control variables and simulating the relationships between these variables and the system through simplified models. In general, the design problem is expressed as the problem of optimizing an objective function under certain constraints; the solution, or solutions, of this problem are then used as starting points for successive extensions and comparative evaluation described in

Chapter 10. Obviously, the more "elementary" the intervention to be designed, the closer the formulation of the supply design model is to the real problem. Thus, the problem of designing traffic-signal control parameters at an isolated intersection can be expressed by an optimization model which, among all the possible values, searches for those minimizing the total delay, or maximizing the total capacity of the intersection. The resulting optimal control parameters can be used directly in the real world. On the other hand, if the problem is to design the transportation system of a region it is practically impossible to represent the complexity of the objectives and constraints. In this case, one or more simplified design models can be formulated, for example, to define the road network, the public transport network and the pricing structure, which jointly or separately minimize the total generalized user costs under budget, technical and environmental constraints. In any case the solution, or the solutions, of the partial problems will only be the starting point for the further phases of design, evaluation and negotiation, which will lead eventually to collective choices. A classification of the design models proposed in the literature and most often used in applications can be made on the basis of some elements described below and summarized in Fig. 9.1.2.

Design (control) variables	*Network topology* *Performances* *Prices and fares*
Objectives	*Social* *Operator's*
Constraints	*External* *Technical* *Demand/flow/cost consistency*
Simulation Model	*Assignment model*

Fig. 9.1.2 Classification of supply design models.

Design (control) variables. Design problems can be divided into three groups with respect to control variables: *network topology*, or layout, (e.g. road network or public transport lines), *performances* of supply elements (e.g. transit line frequencies or traffic-signal control parameters) and *pricing* (e.g. air, railway, parking or motorway fares). The design variables may be discrete (topology and performances) or continuous (prices and performances) according to the specific problem. Obviously a model can, and often does, aim at defining the optimal combination of different types of variables.

Objectives. The design can be developed from different perspectives, i.e. the design model can be defined to optimize (maximizing or minimizing) different objective functions. Design models can account for *social* objectives, e.g. the minimization of total generalized user costs, and/or *operator's* objectives, e.g. the minimization of investment and/or management costs or the maximization of net traffic revenues. The social objectives underlying larger projects are significantly

simplified. The objective function may be *mixed*, i.e. a combination of social and operator's objectives as in Benefits-Costs analysis described in section 10.5.1. Other objective functions correspond to multi-attribute utility functions in multi-objective analysis described in section 10.5.2.

Constraints. Most supply design models can be formulated as constrained optimization problems and, as it is often the case in modeling, some objectives can be introduced as constraints (and vice-versa), for computational convenience. These constraints can be defined *external*, e.g. the maximum available budget, the maximum levels of pollutant concentration. In the former case the implicit objective is to minimize the cost; in the latter, it is to reduce air pollution. *Technical* constraints relate to aspects of the system such as maximum flow-capacity ratios, minimum and maximum frequencies of bus lines, etc. Some specifications of the design model use a third category of constraints representing the consistency between demand, flows, design variables and system performances. These constraints represent the system simulation model and will be considered in the next section.

Simulation model. The simulation model which is usually most relevant to design problems is the *assignment* (or demand-supply interaction) *model.* As shown in previous chapters, such models can be based on within-day static or dynamic system representation, on deterministic or stochastic path choice models and may or may not account for congestion effects. Furthermore, the assignment model may assume *rigid* or *elastic demand* according to whether demand flows are considered constant with respect to the values of the design variables or not.

Although transportation supply design models received considerable scientific and professional attention, they have not reached a level of theoretical completeness and/or number of applications comparable to those described in previous chapters for simulation models. Further, design problems have also not been studied at the same level of detail. It is difficult to present general results for all supply models, as they are specific to the design problem and to a number of assumptions that can be made in connection with each of them. A systematic review of all the supply models presented in the literature and of their transportation engineering implications would require a book on its own. Rather, this chapter will briefly analyze this wide and still open area of application. A general formulation of the supply design models will be described first in section 9.2; some specialization of the general model to the most common design problems will be introduced in section 9.3 without analyzing either the specific models proposed or the implications of related results. Finally section 9.4 describes some algorithms which can be applied to solve various design problems.

9.2. General formulations of the Supply Design Problem

The supply design problem (SDP) can be formulated through a *constrained optimization model*, maximizing or minimizing an objective function $w(.)$, dependent on design variables, y, and link flows, f. The representation of the system and its variables can be within-day static or dynamic. Although some SDP models for dynamic systems are covered in the literature, most of the specifications refer to static systems and assignment models. This is not surprising, given both the recent development and computational complexity of dynamic models, which should be used repeatedly in a SDP. For these reasons, the following will deal with static models. As stated in Chapter 5, link flows resulting from a static assignment model can be expressed as a function of O-D demand flows (vector $\boldsymbol{d}$), of the network topology (link-path incidence matrix Δ) and on path choice probabilities (matrix $\boldsymbol{P}$). In general, both the network topology and path choice probabilities depend on the supply configuration, either directly or through link costs and cost functions. Demand flows are constant if the assignment model assumes rigid demand, and dependent on supply performances if demand is elastic. The *general supply design model* can be formulated as:

$$y^* = \underset{y}{argmax}(\min)\, w(y, f^*) \tag{9.2.1a}$$

subject to the constraints:

$$f^* = \Delta(y)\, \boldsymbol{P}[y, g(f^*, y)]\, d[g(f^*, y)] \tag{9.2.1b}$$

$$y, f^* \in E \tag{9.2.1.c}$$

$$y, f^* \in T \tag{9.2.1.d}$$

where y^* is the optimal solution of supply design problem and f^* the equilibrium flow vector; equation (9.2.1b) expresses the consistency constraint between supply performances, demand and flows (i.e. the equilibrium assignment); equation (9.2.1c) identifies the set of supply parameters satisfying the external constraints and equation (9.2.1d) expresses the system of technical constraints. Furthermore, the notation $\Delta(y)$ indicates that, in the case of design variables influencing the network topology, both the paths and the link-path incidence matrix depend on the values of the design variables; the same is true for path choice probabilities as expressed by $\boldsymbol{P}(y, \boldsymbol{g})$, where $\boldsymbol{g}$ is the path cost vector.

The formulation (9.2.1) is based on the explicit representation of the assignment model with a fixed-point model. As was seen in Chapter 5, this formulation presents some mathematical problems for deterministic user equilibrium (DUE) assignment. In this case, the consistency constraint (9.2.1b) is usually replaced by the variational inequality, which for rigid demand becomes (see section 5.4.2):

$$c(f^*, y)^T (f - f^*) \geq 0 \qquad \forall f \in S(y, d) \tag{9.2.1e}$$

Expression (9.2.1e) makes explicit the dependence of the link flows feasibility set, S, on demand and design parameters. For elastic demand, reference can be made to the analogous expression given in section 5.6.1.2.

The design model can be formulated differently if assignment can be formulated through optimization problems. In this case, model (9.2.1) can be expressed as a *bi-level optimization model* where the value of the first-level objective function, $w(.)$, depends on the solution of a second-level optimization problem, usually with a different objective function, $z(.)$:

$$y^* = \underset{y}{arg\ opt}\ w(y, f(y)) \tag{9.2.2a}$$

$$y \in E$$

$$y \in T$$

$$f(y) = \underset{f \in S_f}{arg\ min}\ z(f, y, d) \tag{9.2.2b}$$

The specific form of the objective function $z(.)$ in (9.2.2b) depends on the specific assignment model (see Chapter 5 for DUE and the SUE specifications). For an uncongested network assignment model, the link costs vector depends exclusively on the design variables, $c = c(y)$, simplifying the specification and the solution of the design model.

The actual specification of the supply design model, whether in the form (9.2.1) or in (9.2.2), comes from the particular design problem and the assumptions. Examples of specifications will be given in the next section. As mentioned earlier, the *design variables* can be divided into three typologies: *topological* or *network layout variables*, usually discrete, denoted in the following with the vector y^{TOP}; *supply performance variables*, continuous or discrete, denoted with y^{PER}; and *price variables*, usually continuous, indicated with y^{PRI}. The vector y of the design variables can therefore, in the most general of cases, be decomposed in the three sub-vectors:

$$y^T = (y^{TOP}, y^{PER}, y^{PRI})^T \tag{9.2.3}$$

The objective function can assume different forms dependent on the goal of the project. *Social objective functions*, $w_1(.)$, usually correspond to the network indicators described in section 5.2. The most common specification is relative to the total actual cost which, in the absence of non-additive path costs, can be expressed as:

$$w_1(\boldsymbol{y},\boldsymbol{f}) = \Sigma_l\, c_l(\boldsymbol{y},\boldsymbol{f})\, f_l \tag{9.2.4}$$

The total Expected Maximum Perceived Utility with respect to path (and possibly mode) choice is seldom adopted as objective function[3] because of its computational complexity, even if it would be a more correct measure of the users' surplus, as will be seen in section 10.4.3:

$$w_1(\boldsymbol{y},\boldsymbol{f}) = \Sigma_{od}\, d_{od}\, s_{od}(-\Delta^T(\boldsymbol{y})\, \boldsymbol{c}(\boldsymbol{y},\boldsymbol{f})) \tag{9.2.5}$$

Operator objective functions, $w_2(.)$, may express the total investment and maintenance cost which depend on design parameters, $\boldsymbol{y}$, or on their functional transformations:

$$w_2(\boldsymbol{y}) = \Sigma_j\, b_j\,(y_j)\, y_j \tag{9.2.6}$$

where b_j is the unitary cost related to each design variable, y_j. For example, if the design variables are zero/one topological variables expressing whether to include or not the connection j, b_j is the investment and/or maintenance cost for that connection. Another type of operators objective function includes the traffic revenues, dependent on the design price variables, which can be either associated with individual links, vector $\boldsymbol{y}_L^{PRI}$, or the O-D pairs, vector $\boldsymbol{y}_{OD}^{PRI}$:

$$w_2\,(\boldsymbol{y}_L^{PRI},\boldsymbol{f}) = \Sigma_l\, y_l^{PRI}\, f_l \tag{9.2.7}$$

$$w_2\,(\boldsymbol{y}_{OD}^{PRI},\boldsymbol{d}) = \Sigma_{od}\, y_{od}^{PRI}\, d_{od} \tag{9.2.8}$$

In the case of *multi-objective optimization*, objective functions are usually expressed as linear combinations of two or more of the above functions. For example, the total cost for the user and for the operator is usually obtained by adding (9.2.4) and (9.2.6) with coefficients representing the relative weight of the two objectives. Furthermore, expression (9.2.6) can also be used to specify an overall (external) budget constraint:

$$\Sigma_j\, b_j(y_j)\, y_j \le B \tag{9.2.9}$$

where B represents the maximum available budget.

Little can be said on the *mathematical properties* of supply design models in general, and on the existence and uniqueness of the solution $\boldsymbol{y}^*$ in particular, since the solution depends on the particular specification adopted. In most cases neither the objective function nor the constraints have convexity properties sufficient for the uniqueness of the solution. In fact, many models have shown multiple solutions, or local optima, corresponding to similar values of the objective function. This may have significant practical implications since nearly equivalent solutions can be generated, among which the best solution can be chosen on the

basis of a wider set of objectives and criteria. Similar considerations can be made for the existence of the solution y^*, which obviously depends on the definition of the constraints; erroneous or incompatible specifications could lead to problems without any feasible solution.

9.3. Some applications of Supply Design models

Supply design models have been studied in greater detail for some classes of "partial" problems, which will be described below. For more complex projects, the actions to be jointly designed may relate to many elements of the supply system and to many modes. In the case of tactical urban transportation planning, for example, actions may include the directions and traffic-signal control of the road network, the availability of parking areas on and off-street, the structure and frequency of the transit lines, parking and transit pricing, and so on. Similarly, for a railway system program, design variables may include the structure of the lines, the timetables of individual runs and the fare structure. Design problems of this complexity are usually solved by formulating separate design models for one or more individual components following a sequence related to the (implicit) hierarchy of the objectives.

9.3.1. Models for road network layout design

Design problems in this class identify the road connections to be built or the optimal circulation scheme for a given network of facilities. The design variables for these models are discrete topological variables represented by the vector y^{TOP} with a component for each possible road connection. These variables are a subset of the expanded road network links that include the existing connections as well as the possible connections to be designed.

Typically in the *optimal infrastructure layout problem*, roads are assumed to be bi-directional and the design variables are usually binary variables $y_j^{TOP} = 0/1$, indicating that the link j is to be excluded (zero) or included (one) in the solution. Fig. 9.3.1 shows an example of the initial configuration and some possible alternative configurations with the relative values of the design variables for a small test network. This SDP has been often associated with extra-urban road networks.

Usually, the objective function is specified as a linear combination of the total transportation cost (9.2.4) and the total construction and maintenance cost (9.2.6) where $b_j(y_j^{TOP})$ is the cost to build and operate the road connection represented by link j. To ensure comparability of the two terms, transportation and construction/operation costs should be expressed in monetary units and cover the same period, e.g. a generic average year. This can be accomplished by "projecting" into a given year (typically the first year of operation) the values of O-D flows and the user annual transportation costs. Similarly, the operator cost will be the equivalent yearly amount of the total investment cost and the yearly maintenance cost.

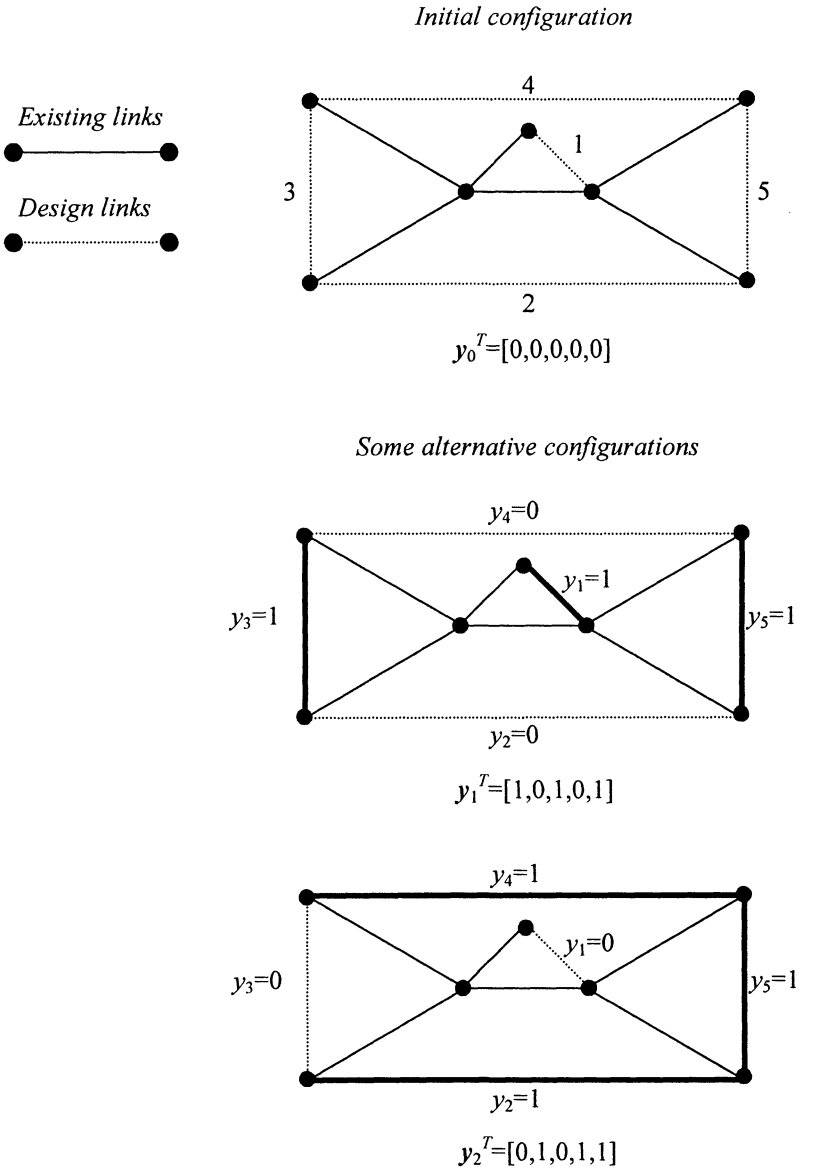


Fig. 9.3.1 Design variables for an optimal infrastructures layout problem.

External constraints usually include a budget constraint (9.2.9) and, in some cases, constraints on the total level of pollutants emitted. Some specifications may include the so-called "network constraints" ensuring the connection of all Origin-Destination pairs, the node conservation of flows, etc., as described in section 5.2. It should be noted, however, that network constraints are necessary only if the assignment model is deterministic (DUN or DUE) and is expressed by a variational inequality or an optimization model in terms of link variables. In fact, in stochastic assignment models, these constraints are implicit in the relationships between demand and link flows as expressed by eqn. (9.2.1b). Many specifications of this model consider rigid demand, using the modal O-D matrices forecasted for the reference year.

A simplified specification of the design problem is:

$$\boldsymbol{y}^{*TOP} = \underset{\boldsymbol{y}^{TOP}}{argmin}\ \Sigma_l c_l\left(\boldsymbol{y}^{TOP}, \boldsymbol{f}^*\right)\cdot f_l^* \tag{9.3.1}$$

subject to the constraints:

$$y_j^{TOP} = 0/1$$

$$\boldsymbol{f}^* = \Delta(\boldsymbol{y}^{TOP})\ \boldsymbol{P}[\boldsymbol{y}^{TOP}, \boldsymbol{g}(\boldsymbol{f}^*, \boldsymbol{y}^{TOP})]\ \boldsymbol{d}$$

$$\Sigma_j\, y_j^{TOP}\cdot b_j \le B$$

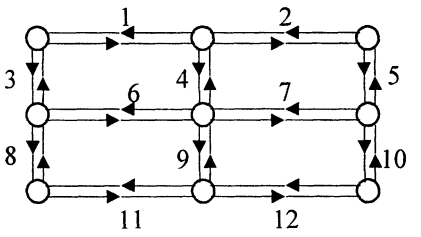
The *optimal functional layout problem* considers the optimal circulation scheme, i.e. the optimal configuration of traffic directions for a road network, typically an urban network. The need for optimal circulation schemes arises for two conflicting reasons. The single-direction use of a road increases the available width for this direction and, in turn, the saturation flow at the final intersection. This reduces the waiting time for a given flow. On the other hand, two-way roads generally reduce the distance between an O-D pair and increase the conflict points at the intersections. The design variables are still discrete variables, $\boldsymbol{y}^{TOP}$, associated with each link and can assume different values (e.g. 0, 1, 2) according to whether the link is used in both directions or in each of the two ways (see Fig. 9.3.2).

The cost functions of each link j depend on the variable y_j; furthermore, the objective function usually includes only the user generalized cost (9.2.4), the construction cost of existing roads being null and the difference in operation costs being negligible. The link constraints are analogous to those described for the infrastructures layout problem and the same considerations hold.

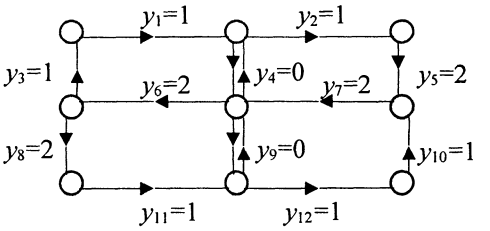
The model is sometimes specified by introducing external constraints, limiting, in particular, flow/capacity ratios for relevant links. This constraint expresses the need for both technical functionality (flows near capacity induce instability phenomena and possible spill-backs at intersections) and pollution reduction

(emissions are higher for low commercial or average speeds). Another type of external constraint requires that the distance between each O-D pair on the shortest path not exceed the "shortest" feasible distance, i.e. the minimum distance on a fully bi-directional road configuration, by more than a prefixed amount. In this case the implicit "equity" objective is to distribute penalties among users.

$y_0^T = [0,0,0,0,0,0,0,0,0,0,0,0]$



$y_1^T = [1,1,1,0,2,2,2,2,0,1,1,1]$



$y_2^T = [0,0,0,1,0,2,2,0,1,0,0,0]$

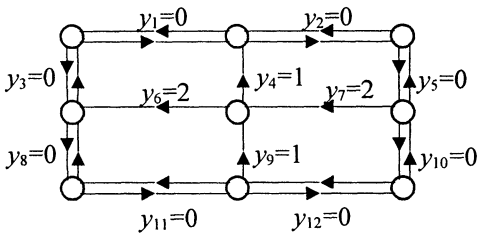


Fig. 9.3.2 Design variables for an optimal functional layout problem.

The optimal urban road network layout problem, discussed below, is usually associated with the control of intersections that determine road link capacity. As was seen in section 2.3.1.2., the capacity of a signalized road access is given by the product of the saturation flow by effective green to cycle length ratio.

9.3.2. Models for road network capacity design

Capacity design models optimize the capacity of links in a road network of given topology. The design variables are generally continuous, expressed by a vector y^{PER} whose components are link capacities. The problem may assume two different forms, typical of extra-urban and urban road networks.

For *extra-urban road network capacity design*, the decision variables are the link capacities, usually constrained by pre-set minimum and maximum values. The formulation of the model is substantially similar to that described for the optimal network layout problem; the objective function can be expressed as a sum of the user costs and the construction costs. Budget and congestion-level constraints (maximum value of the flow/capacity ratios) are also typically included.

The capacity of an extra-urban road depends on first approximation on its transversal section (number of carriageways and their width, lateral distances, etc.) and does not assume all the possible values, but only some discrete values corresponding to the different section typologies. From this point of view, the design variables should be discrete even though, in the literature, they are often approximates as continuous variables. In the other case (discrete capacity), the problem would be analogous to the one described in the previous section, with the difference that the design variables can assume several discrete values corresponding to the different section typologies.

The *urban road network capacity design* often addresses the problem of finding optimal traffic-signal control parameters for a subset of intersections (*traffic signal setting problem*). In the most simplified formulations, it is usually assumed that intersections are "isolated", i.e. the traffic-signal coordination between adjacent intersections has no effect. This assumption implies that *offsets* between the green times of different intersections are not relevant control variables. It can also be assumed that for each intersection the overall duration of the cycle and the structure of the traffic-signal phases are known. This implies that for each node (or group of nodes), n, representative of a signalized intersection, the set J_n of the phases j_n and the set of links $I(j_n)$ corresponding to flows receiving green in the same phase is known. In this case, the design variables y^{PER} can be identified as the effective green to cycle length ratios, the latter deducted of the lost times for each phase. The design variables are therefore continuous over the interval 0, 1 (see Fig. 9.3.3).

Note the difference between capacity design of signalized intersections for the entire network and for a single intersection. In the first case, as the green/cycle ratios vary, capacities vary, and because of the effect of assignment constraints (9.2.1b), link flows vary. In the case of a single intersection, it is assumed that flows are known and invariant with respect to capacity parameters.

The specification usually adopted for the design model is analogous to that described for the road network lay-out problem. The objective function to be minimized is the total generalized cost (usually time) spent on the network. Construction costs are not taken into consideration and the external constraints

might include maximum levels of congestion and pollution. Technical constraints set the maximum and minimum duration of each phase and require that the summation of green/cycle ratios over all phases is equal to one for each intersection.

Two different approaches can be followed to optimize the traffic-signal control parameters: *local* and *global*. In the *global* approach the control parameters of all intersections are jointly optimized to minimize the total travel time on the network. In the *local* approach, each signalized intersection is optimized to minimize the total user delay at the intersection. In this case, a circular dependence between flows, costs and control parameters arises and the resulting problem can be seen as a *fixed-point* problem. This problem can be modeled as an asymmetric user assignment problem.

A possible simplified formulation of the global optimal signal setting problem is:

$$y^{*PER} = \underset{y^{PER}}{argmin}\, \Sigma_l c_l\left(y^{PER}, f^*\right)\cdot f_l^* \tag{9.3.2}$$

subject to:

$$0 \le y_{jn}^{PER} \le 1$$

$$\sum_{jn \in J_n} y_{jn}^{PER} = 1 \qquad \forall n$$

$$y_{jn}^{PER}\, T_{c\,n} \ge T_{min} \quad \forall n$$

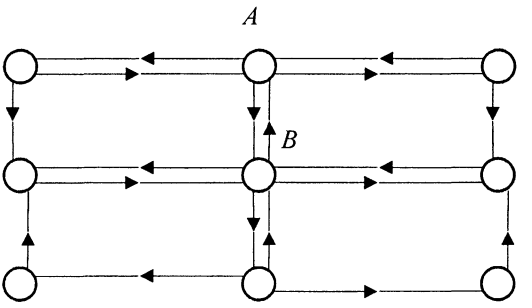
$$f^* = \Delta\, \boldsymbol{P}[y^{PER}, g(f^*, y^{PER})]\, \boldsymbol{d}$$

where $T_{c\,n}$ is the duration of the cycle at intersection n and T_{min} is the minimum value for a green time interval.

More complex traffic-signal control problems introduce other design variables and in particular off-sets between green times in nearby intersections, cycle length for each intersection and sequence and number of phases. In the first case link delay models described in Chapter 2 must account for the effects of platoon dispersion between coordinated intersections.

9.3.3. Models for transit network design

It is usually assumed that the relevant supply variables for high frequency urban transit systems are service frequencies rather than actual timetables (see section 2.5). Under this assumption the design problem identifies the optimal layout for the lines and their service frequencies in the reference period (e.g. rush hour). In this case the design model simultaneously identifies the topological configuration and the optimal performances of a supply system.



Intersection A

Cycle T_{cA}

Total lost time LT_A

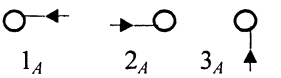
Phases $J_A = \{1_A, 2_A, 3_A\}$

Variables $y^{PER}_{1A} = G_{1A}/(T_{cA} - LT_A)$

$y^{PER}_{2A} = G_{2A}/(T_{cA} - LT_A)$

$y^{PER}_{3A} = G_{3A}/(T_{cA} - LT_A)$

Phase plan



Constraint on total cycle length

$G_{1A} + G_{2A} + G_{3A} = T_{cA} - LT_A$

Intersection B

Cycle T_{cB}

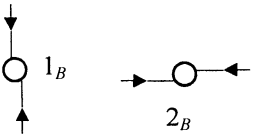
Total lost time LT_B

Phases $J_B = \{1_B, 2_B\}$

Variables $y^{PER}_{1B} = G_{1B}/(T_{cB} - LT_B)$

$y^{PER}_{2B} = G_{2B}/(T_{cB} - LT_B)$

Phase plan



Constraint on total cycle length

$G_{1B} + G_{2B} = T_{cB} - LT_B$

Vector of variables

$y^{PER} = [y^{PER}_{1A}, y^{PER}_{2A}, y^{PER}_{3A}, y^{PER}_{1B}, y^{PER}_{2B}]^T$

Constraints

$y^{PER}_{1A} + y^{PER}_{2A} + y^{PER}_{3A} = 1$

$y^{PER}_{1B} + y^{PER}_{2B} = 1$

Fig. 9.3.3 Design variables and constraints for an optimal signal setting problem.

The design variables are the discrete layout variables, y^{TOP}_{ln}, equal to one if the physical link l, e.g. road or railway section, belongs to the line n, and zero otherwise), and the continuous performance variables, y^{PER}_{n}, representing the service frequency of each line n, see Fig. 9.3.4. The lay-out variables are equivalent to the duplication of physical links in line links i.e. to the implicit construction of the line network model described in Chapter 2. For this reason, the link variables of this model will be expressed with the double index.

The objective function usually includes the user generalized cost and the operator cost, appropriately homogenized. For urban transit systems the functions $w_1(.)$ expressing user costs is different from (9.2.4) given the usual assumptions on mixed preventive/adaptive path choice behavior. In this case alternative travel strategies are represented by hyperpaths on the lines, and average path costs include a non-additive component associated with waiting times at stops (see section 4.3.4).

Formally, the objective function $w_1(.)$ can be expressed as:

$$w_1(y^{TOP}, y^{PER}) = \Sigma_n \Sigma_{l \notin Jw} c_{ln} f_{ln}(y^{TOP}, y^{PER}) + \Sigma_{l \in Jw} \Sigma_k tw^k_l (y^{TOP}, y^{PER}) f^k_l (y^{TOP}, y^{PER}) \quad (9.3.3)$$

where c_{ln} is the generic additive cost associated with the link l and with line n (e.g. on-board or access travel times); J_w is the set of waiting links, k the generic hyperpath and tw^k_l the waiting time (cost) associated, with the link l and with the hyperpath k, f_{ln} is the users' flow on on-board link l belonging to line n and f_{lk} is the flow on waiting link l belonging to hyperpath k, see section 4.3.4.

The overall operator cost $w_2(.)$ is usually expressed using the unit running cost CE_n for each journey (bus, train, etc.) of the line n, expressed in monetary units per distance or time unit:

$$w_2(y^{TOP}, y^{PER}) = \Sigma_n \Sigma_l y^{TOP}_{ln} CE_n L_{ln} y^{PER}_{n} \quad (9.3.4)$$

where L_{ln} is the length (or round time) of the link l for the line n.

The assignment constraints can be expressed using the formulation introduced in Chapter 5, as:

$$\boldsymbol{f}(y^{TOP}, y^{PER}) = \boldsymbol{\Lambda}(y^{TOP}, y^{PER})\, \boldsymbol{Q}(\boldsymbol{\Lambda}^T(y^{TOP}, y^{PER})\, \boldsymbol{c}(y^{TOP}) + \boldsymbol{x}^{NA}(y^{TOP}, y^{PER}))\, \boldsymbol{d} \quad (9.3.5)$$

where it is implicitly assumed that the network is not congested and that the link crossing probability matrix $\boldsymbol{\Lambda}$, and the non-additive hyperpath costs, $\boldsymbol{x}^{NA}$, both depend on the topological configuration of the lines (y^{TOP}) and on the respective frequencies (y^{PER}).

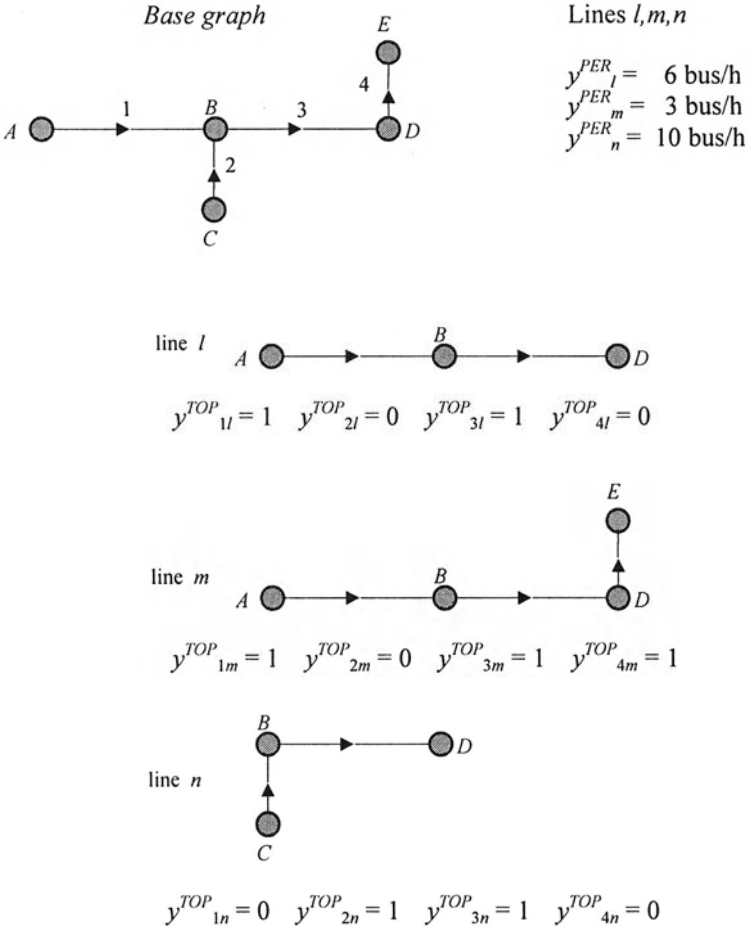


Fig. 9.3.4 Design variables for an optimal line layout and frequency problem.

The technical constraints of the problem usually restrict the flow f_{ln} on each line link l to the capacity of the line n, which can be expressed as the product of the capacity Cap_n of each vehicle and the frequency of the line n:

$$f_{ln} \leq Cap_n y^{PER}_n \qquad \forall\, l$$

Furthermore, the frequencies must be non-negative, equal to zero if the line is not active and below a maximum technically feasible value y^{PER}_{max}:

$$0 \leq y^{PER}_n \leq y^{PER}_{max}$$

Another possible technical constraint is a budget constraint on the vehicle stock. This constraint can be expressed as a function of the travel time of each line link, t_l, since the number of journeys necessary for a line of frequency $y^{PER}{}_n$ is equal to the product of the frequency for the total travel time of the line:

$$\Sigma_n \Sigma_l t_l y^{PER}{}_n \cdot y^{TOP}{}_{ln} \leq N_{max} \tag{9.3.6}$$

with N_{max} equal to the maximum number of available vehicles.

Finally, a technical constraint sometimes introduced, though not easily expressed in formal terms, requires that lines must have their terminals in a given set of nodes.

A simplified version of the transit design problem assumes the topological configuration of all the lines, components of the vector y^{TOP}, as given. In this case the design problem is reduced to the calculation of optimal service frequencies, i.e. the components of the vector y^{PER}, with a significant reduction in the number of variables and in computational complexity. A minimal service frequency may be added to the technical constraints.

For *extra-urban services* (low frequency, high regularity) the supply design problem is quite different, as are the models used to simulate these services. As was seen in Chapter 6, the diachronic network models used to simulate regular low frequency services are based on the explicit representation of the service-schedule. Optimal scheduling design models define departing and arrival times of each run of a pre-defined set of lines. Furthermore, in the most general case, they jointly determine the lines and their departure and arrival times, under a set of technical constraints. The latter are the feasible range of travel times (feasible commercial speeds), the available vehicle stock, the range of acceptable connection times between different lines at intermediate stops, etc. The problem of optimal service scheduling has not been covered extensively in the literature.

9.3.4. Models for pricing design

Pricing design models can be applied to different contexts. Prices, generally represented as continuous variables y^{PRI}, may be related to the different transportation supply elements: road tolls, parking, air and railway fares, etc. The specification of the design variables y^{PRI} will depend on the assumed "pricing structure", i.e. on how the prices are computed and applied. If constant access prices are assumed, e.g. constant road tolls at motorway entrance/exit points or constant parking fares, the components $y^{PRI}{}_j$ of the vector can be associated with the network links, j, representative of the toll points or of the parking facilities. If the price is proportional to the distance covered, e.g. road tolls or railway fares proportional to the journey length, the price parameter $y^{PRI}{}_l$ can be associated to each link, l, corresponding to a section with a physical length.

The objectives of pricing design might also be different. If the pricing policy is meant to improve the efficiency of the transportation system, for example by

reducing the overall generalized cost of the system and/or the overall pollution level, the resulting pricing is known as *efficiency pricing*. A typical example of efficiency pricing design is *road pricing*, i.e. the application of a price to roads to minimize the non-renewable total cost, typically the total travel time. In this case the social objective function[4] is $w_1(y^{PRI}) = \Sigma_l\, c_l\,(f_l)\, f_l$. The efficiency road pricing design problem whit rigid demand can therefore be formalized as:

$$y^{*PRI} = \underset{y^{PRI}}{argmin}\ \Sigma_l\, c_l(\boldsymbol{f}^*)\, f_l^* \tag{9.3.7}$$

subject to the constraints:

$$y^{PRI} \geq \mathbf{0}$$

$$\boldsymbol{f}^* = \Delta\, \boldsymbol{P}(\boldsymbol{g}(\boldsymbol{f}^*, y^{PRI}))\, \boldsymbol{d}$$

In the special case of DUE assignment with separable cost functions, the problem (9.3.7) is equivalent to the system optimal (SO) assignment problem described in section 5.4.4. This problem is a single-level optimization problem and the optimum price y_l^{*PRI} can be calculated as:

$$y_l^{*PRI} = c'_l\,(f_l^*)\, f_l^* \qquad \forall\, l \tag{9.3.8}$$

where $c'_l(f_l)$ is the first derivative of the cost function.

However, it should be noted that the prices vector y^{*PRI} given by (9.3.8), is not the unique solution to the problem (9.3.7) in general. Under deterministic path choice, there may be other vector solutions to the general problem (9.3.7) with different operational impacts (e.g. less expense to the users or the possibility of applying the price only to certain network links).

The formulation (9.3.7) of the road pricing problem assumes that O-D demand $\boldsymbol{d}$ is rigid. The resulting prices tend to reduce the total travel time by modifying path choices; this is achieved by increasing generalized link costs proportionally to their congestion levels. However, many empirical results indicate that the most significant congestion reductions can be obtained by focusing on demand flows. To address this problem it is necessary to consider the O-D demand for the car mode, $\boldsymbol{d}^C$, as elastic. For example, it may be assumed that demand is elastic with respect to modal choice and that road is the only congested system, i.e. that only road costs are dependent on link flow $\boldsymbol{f}^C$ and design prices y^{*PRI}. It may be appropriate to impose further constraints to the problem (9.3.7), for example requiring that the road link flows are below a predetermined fraction of the corresponding capacities.

Under these assumptions the efficiency road-pricing problem with elastic demand model can be reformulated as:

$$\boldsymbol{y}^{*PRI} = \underset{\boldsymbol{y}^{PRI}}{\arg\min} \sum_l c_l(\boldsymbol{f}^*) f_l^* \tag{9.3.8}$$

subject to:

$$\boldsymbol{y}^{PRI} \geq \boldsymbol{0}$$

$$\boldsymbol{f}^* = \Delta \boldsymbol{P}(\boldsymbol{g}^C(\boldsymbol{y}^{PRI}, \boldsymbol{f}^*))\, \boldsymbol{d}^C(\boldsymbol{s}^C(\boldsymbol{y}^{PRI}, \boldsymbol{f}^*)\, \boldsymbol{s}^B)$$

where $\boldsymbol{s}^C(.)$ and $\boldsymbol{s}^B$ are $(n_{od} \times 1)$ vectors of EMPU variables related to path choice for car and bus modes.

The pricing design model has a different form when maximizing traffic revenues or net profits (revenues minus costs). In this case, assuming a single operator in the market, the operator's objective function $w_2(.)$ is the total revenue, and the problem can be formulated as:

$$\boldsymbol{y}^{*PRI} = \underset{\boldsymbol{y}^{PRI}}{argmax}\ \Sigma_l\, \boldsymbol{y}^{PRI}{}_l f_l^* \tag{9.3.9}$$

subject to:

$$\boldsymbol{f}^* = \Delta \boldsymbol{P}(\boldsymbol{g}^C(\boldsymbol{y}^{PRI}, \boldsymbol{f}^*))\, \boldsymbol{d}^C(\boldsymbol{g}^C(\boldsymbol{y}^{PRI}, \boldsymbol{f}^*), \boldsymbol{g}^B))$$

$$\boldsymbol{y}^{PRI} \geq \boldsymbol{0}$$

where the O-D demand flows relative to mode is considered price elastic for non-marginal variations of the latter.

Pricing design models for other transport infrastructures (e.g. rail lines or airport slots) and services (e.g. train or air connections) can be formulated in a similar way. Typically optimal infrastructure use prices are computed with respect to social objectives (efficiency pricing) while service prices are computed with respect to operators objectives. Limited applications of these models are described in the literature.

9.3.5. Models for mixed design

Complex projects involving an area-wide transportation system or several aspects of the services provided by a company would require design models integrating two or more of the models described earlier. For example, a regional transportation plan usually includes the optimal design of road and rail infrastructures, rail and bus services, road and transit pricing and so on. Similarly the definition of a road project-financing scheme includes the optimal design of new infrastructures and pricing systems. Clearly the computational complexity of these problems increases exponentially and the (few) examples published in the literature are based on a number of ad hoc simplifying hypotheses specific to the individual problem. Solution algorithms generally are based on the sequential solution of separate design problems corresponding to separate design variables.

9.4. Some algorithms for Supply Design Models

Using mathematical programming terminology, supply design models can be specified as discrete, continuous or mixed optimization problems; such models are generally non-linear with non-linear constraints, or bi-level optimization models with ill-defined mathematical properties. For most of these problems, there are not optimal algorithms, i.e. algorithms, which can be proven to converge towards global or local optimal solutions.

For these reasons, heuristic algorithms have been used in applications, which in many cases provide satisfactory results. This is especially relevant considering that the goal is to define actions about the physical system with the help of design models rather than to solve a mathematical problem per se. In what follows, some examples of heuristic algorithms will briefly be presented for discrete and continuous problems. These algorithms are applicable to a wide range of design models. A comprehensive review is beyond the scope of this book.

9.4.1. Algorithms for the discrete SDP

Several algorithms have been proposed for solving discrete SDP; most solve specific network design problems. It is possible to classify these algorithms in two groups:

- *Exact algorithms* that yield optimal solutions (global optimum), such as total enumeration and "branch and bound" algorithms.
- *Heuristic algorithms* that yield sub-optimal solutions (local optimum or near optimal solution), such as add-and-delete algorithms, neighborhood search algorithms, genetic algorithms, simulated annealing algorithms.

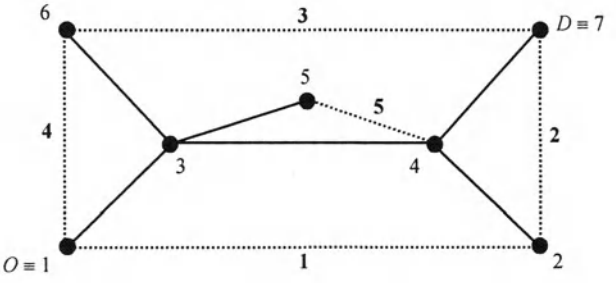
In general, exact algorithms can be applied only to small networks, while heuristic algorithms can be applied to relatively large networks. For comparison sake, algorithms will be applied to the small network of Fig. 9.4.1, for the uncongested road layout design problem with deterministic route choice model. The algorithms can be extended easily to other discrete design problems (e.g. optimal transit lines layout).

Add-and-delete algorithms. They perform a sequence of insertion and deletion routines starting from an initial solution. The insertion routine adds design links sequentially to generate new solutions. For each possible addition, the objective function value is calculated and the link with the largest objective function improvement is added to the current configuration. The routine continues to add links until no link insertion improves the objective function. The deletion routine deletes links from the current configuration, calculating the objective function with each deletion. The link with the largest improvement of objective function is deleted from the current configuration. The routine continues to delete links until no link deletion improves the objective function. If at least one link is deleted, the

algorithm repeats the insertion; otherwise the algorithm stops. In the last generated network configuration no link insertion or deletion could improve the objective function value.

The vector of design variables at iteration *it*, including or excluding link *l* is denoted $y_{it,l}^{TOP}$. The results of an application of the *add-and-delete* algorithm to the test network of Fig. 9.4.1 are summarized in Fig. 9.4.2.

In order to accelerate calculations, sequentially inserted links can be saved as long as they improve the objective function. Similarly the delete routine may process each link from a list eliminating a link each time the objective function improves.



——— *existing link*

..... *design link*

Objective function:
$w = \boldsymbol{c}^T f(\boldsymbol{y}^{TOP}, \boldsymbol{d}) + \beta \boldsymbol{b}^T \boldsymbol{y}^{TOP}$
$\beta = 5$ time units/cost unit

Link number	*Link*	*Travel time* (**c**)	*Building cost* (**b**)
	Design links		
1	1-2	8	800
2	2-7	12	1000
3	6-7	15	1500
4	1-6	10	600
5	4-5	20	500
	Existing links		
	1-3	10	-
	2-4	10	-
	3-4	15	-
	3-5	10	-
	3-6	10	-
	4-7	10	-

Demand OD ($d_{1\text{-}7}$)	1000

Fig. 9.4.1 Test network (starting configuration).

Solution	User costs	Construction costs	Objective function value
Starting solution			
y_0^{TOP} = [0,0,0,0,0]	35000	0	35000
Insertion routine			
First insertion			
$y_{1,1}^{TOP}$ = [**1**,0,0,0,0]	28000	4000	32000
$y_{1,2}^{TOP}$ = [0,**1**,0,0,0]	35000	5000	40000
$y_{1,3}^{TOP}$ = [0,0,**1**,0,0]	35000	7500	42500
$y_{1,4}^{TOP}$ = [0,0,0,**1**,0]	35000	3000	38000
$y_{1,5}^{TOP}$ = [0,0,0,0,**1**]	35000	2500	37500
Best inserted link			
y_1^{TOP} = [1,0,0,0,0]	28000	4000	32000
Second insertion			
$y_{2,2}^{TOP}$ = [1,**1**,0,0,0]	20000	9000	29000
$y_{2,3}^{TOP}$ = [1,0,**1**,0,0]	28000	11500	39500
$y_{2,4}^{TOP}$ = [1,0,0,**1**,0]	28000	7000	35000
$y_{2,5}^{TOP}$ = [1,0,0,0,**1**]	28000	6500	34500
Best inserted link			
y_2^{TOP} = [1,1,0,0,0]	20000	9000	29000
Third insertion			
$y_{3,3}^{TOP}$ = [1,1,**1**,0,0]	20000	16500	36500
$y_{3,4}^{TOP}$ = [1,1,0,**1**,0]	20000	12000	32000
$y_{3,5}^{TOP}$ = [1,1,0,0,**1**]	20000	11500	31500
No inserted link improves objective function			
Deletion routine			
First deletion			
$y_{3,1}^{TOP}$ = [**0**,1,0,0,0]	35000	5000	40000
$y_{3,2}^{TOP}$ = [1,**0**,0,0,0]	28000	4000	32000
No deleted link improves objective function			
Optimal solution			
y_{opt}^{TOP} = [1,1,0,0,0]	20000	9000	29000

Fig. 9.4.2 Add and delete algorithm applied to the test network of Fig. 9.4.1.

Neighborhood search algorithms. These algorithms, starting with an initial solution, generate the set of solutions, which can be reached directly from the current solution by an elementary operation, called *move*. This solution is named a *neighbor* of the current solution and the set of all neighbors is named *Neighborhood.* Among all neighbors the next solution is chosen, selecting either the optimal solution (descent/ascent method) or a random solution (Monte Carlo method). The algorithm ends when no neighbor for the current solution improves

the objective function in the descent method, and when the objective function does not significantly improve over the last m iterations in the Monte Carlo method. The results of an application of the *neighborhood search* algorithm to the test network of Fig. 9.4.1, using the descent method, are summarized in Fig. 9.4.3.

Solution	User costs	Construction costs	Objective function
Starting solution			
y_0^{TOP} = [0,0,0,0,0]	35000	0	35000
Neighborhood generation			
$y_{1,1}^{TOP}$ = [**1**,0,0,0,0]	28000	4000	32000
$y_{1,2}^{TOP}$ = [0,**1**,0,0,0]	35000	5000	40000
$y_{1,3}^{TOP}$ = [0,0,**1**,0,0]	35000	7500	42500
$y_{1,4}^{TOP}$ = [0,0,0,**1**,0]	35000	3000	38000
$y_{1,5}^{TOP}$ = [0,0,0,0,**1**]	35000	2500	37500
Next solution			
y_1^{TOP} = [1,0,0,0,0]	28000	4000	32000
Neighborhood generation			
$y_{2,1}^{TOP}$ = [**0**,0,0,0,0]	35000	0	35000
$y_{2,2}^{TOP}$ = [1,**1**,0,0,0]	20000	9000	29000
$y_{2,3}^{TOP}$ = [1,0,**1**,0,0]	28000	11500	39500
$y_{2,4}^{TOP}$ = [1,0,0,**1**,0]	28000	7000	35000
$y_{2,5}^{TOP}$ = [1,0,0,0,**1**]	28000	6500	34500
Next solution			
y_2^{TOP} = [1,1,0,0,0]	20000	9000	29000
Neighborhood generation			
$y_{3,1}^{TOP}$ = [**0**,1,0,0,0]	35000	5000	40000
$y_{3,2}^{TOP}$ = [1,**0**,0,0,0]	28000	4000	32000
$y_{3,3}^{TOP}$ = [1,1,**1**,0,0]	20000	16500	36500
$y_{3,4}^{TOP}$ = [1,1,0,**1**,0]	20000	12000	32000
$y_{3,5}^{TOP}$ = [1,1,0,0,**1**]	20000	11500	31500
No neighbor improves objective function			
Optimal solution			
y_{opt}^{TOP} = [1,1,0,0,0]	20000	9000	29000

Fig. 9.4.3. Neighborhood search algorithm applied to the test network of Fig. 9.4.1.

The neighborhood search algorithm is similar to the add-and-delete algorithm described previously; the main difference is the sequence in which insertions and deletions are performed. In add-and-delete algorithms a link can be added to the current solution only in insertion routines and can be deleted only in deletion routines. In the neighborhood search, at each step all the links can be added or deleted. However, some applications show that neighborhood search algorithms are better suited to find local optima close to the starting solution and, thus, should

be used as second-step algorithms coupled with other algorithms spanning the whole feasible set.

Genetic algorithms. These algorithms, used for combinatorial problems, mimic the mechanics of genetic and natural selection. These heuristic algorithms, starting with a *population* (set of initial feasible solutions), iteratively generate a new population with a higher probability of containing the optimal (or sub-optimal) solution. Each feasible solution is an *element* (named *chromosome*) of the population, composed of *genes.* A gene is a group of variables satisfying "local" constraints such as the number of lanes to be allocated in each direction for an urban road network design problem. Future populations are generated with three routines: *reproduction*, *crossover* and *mutation.*

The *reproduction* routine generates a new population randomly so that the solutions with higher values of the objective function have higher probability to survive; in this way only the fitter solutions will be submitted to crossover and mutation routines. Survival probabilities are defined by the *fitness function, a* monotone increasing (decreasing) function of the objective function for maximization (minimization) problems. One possible specification of the fitness function is:

$$ff(i) = exp(-\alpha w_i)$$

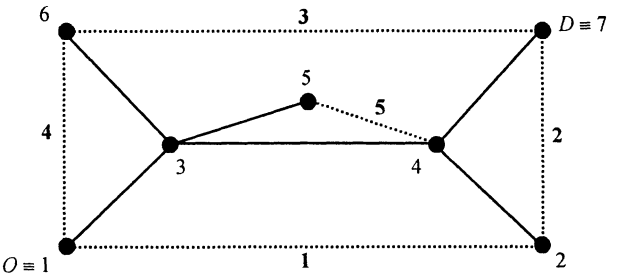
where i is the generic element of the current population (a feasible solution), α is a parameter and w_i is the corresponding value of the objective function. Reproduction probabilities can be computed as:

$$p_r(i) = \frac{ff(i)}{\sum_j ff(j)} = \frac{exp(-\alpha w_i)}{\sum_j exp(-\alpha w_j)}$$

where the summation is extended to all the elements of the current population. The *crossover* routine generates a new population by randomly exchanging parts (genes) between the feasible solutions (chromosomes). The *mutation* routine generates a new population by randomly "mutating" a gene (variable) of a "chromosome" (solution). The algorithm stops when the objective function no longer improves with new solutions, e.g. in their average or min/max values, over the last iterations.

One of the differences of genetic algorithms with respect to the previous ones is that the outcome of the former is a population of feasible solutions, with similar values of the mono-dimensional objective function. Comparisons can be made among these values on the basis of the individual components of the objective function as well as of other variables. Vice versa, optimization algorithms are intended to give a single "best" solution.

The algorithm can be adjusted by setting the parameters of the fitness functions, as well as the number of crossover and mutation routines, at each iteration. An example of a cycle of reproduction - crossover - mutation is reported in Fig. 9.4.4 for a road network design problem to determine the number of lanes in each direction. In this case each gene represents the configuration of a given road and has two components, one for each direction.



——— *existing link*

..... *design link*

Objective function:
$w = \boldsymbol{c}^T \boldsymbol{f}(\boldsymbol{y}^{TOP}, \boldsymbol{d}) + \beta \boldsymbol{b}^T \boldsymbol{y}^{TOP}$
$\beta = 5$ time units/cost unit

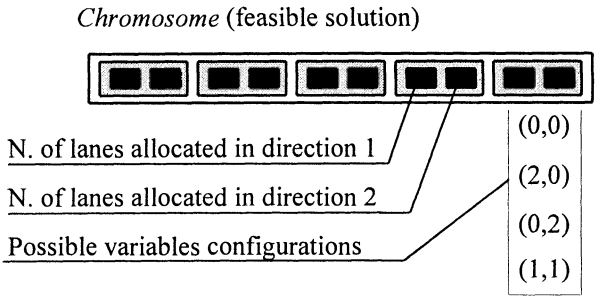


Fig. 9.4.4a Genetic algorithm for a discrete road network design problem: test network.

Parameters of the algorithm	
Number of design links	5
Design variables	lanes in each direction
Lanes in each design link	2
Population	3 elements
Fitness Function	$\exp(-0.0001\ w_i)$
Number of Crossover	1
Number of Mutation	1

Starting Population		Objective function
Solution 1: Present configuration	y_1 = [0,0; 0,0; 0,0; 0,0; 0,0]	35000
Solution 2: Random configuration 1	y_2 = [1,1; 1,1; 1,1; 0,0; 0,0]	44000
Solution 3: Random configuration 2	y_3 = [0,0; 0,0; 1,1; 1,1; 1,1]	46500
Reproduction		
Reproduction Probability (RP)	$\exp(-0.0001\ w_i)/\sum_j \exp(-0.0001\ w_j)$	
RP1 = 58.8 %	range [0; 0.588)	
RP2 = 23.5 %	range [0.588; 0.823)	
RP3 = 17.7 %	range [0.823; 1]	
Random number extraction	New population	
0.456 → Solution 1	y_1 = [0,0; 0,0; 0,0; 0,0; 0,0]	35000
0.672 → Solution 3	y_2 = [0,0; 0,0; 1,1; 1,1; 1,1]	46500
0.089 → Solution 1	y_3 = [0,0; 0,0; 0,0; 0,0; 0,0]	35000
Crossover		
Random solution selection		
Solution 1	y_1 = [0,0; 0,0; 0,0; 0,0; 0,0]	
Solution 2	y_2 = [0,0; 0,0; 1,1; 1,1; 1,1]	
Random cut points selection		
Point 1 → 2	[x,x; x,x; | x,x; x,x; x,x]	
Point 2 → 4	[x,x; x,x; x,x; x,x; | x,x]	
New population		
Solution 1 (crossed)	y_1 = [0,0; 0,0; | 1,1; 1,1; | 0,0]	44000
Solution 2 (crossed)	y_2 = [0,0; 0,0; | 0,0; 0,0; | 1,1]	37500
Solution 3	y_3 = [0,0; 0,0; 0,0; 0,0; 0,0]	35000
Mutation		
Random solution selection		
Solution 3	y_3 = [0,0; 0,0; 0,0; 0,0; 0,0]	
Random mutation link	Link 1 → [**x,x**; x,x; x,x; x,x; x,x]	
New random link configuration	y_3 = [2,0; 0,0; 0,0; 0,0; 0,0]	32000
New population		
Solution 1	y_1 = [0,0; 0,0; 1,1; 1,1; 0,0]	44000
Solution 2	y_2 = [0,0; 0,0; 0,0; 0,0; 1,1]	37500
Solution 3 (Mutated)	y_3 = [2,0; 0,0; 0,0; 0,0; 0,0]	32000

Fig. 9.4.4b Genetic algorithm for the discrete road network design problem of Fig. 9.4.4a.

9.4.2. Algorithms for the continuous SDP

The algorithms for continuous supply design problems are based on the principles of *nonlinear optimization* (see Appendix A). The optimal solution can be expressed in a closed form only for few simple problems (e.g. transit frequency optimization for a single line or cycle length, and green/cycle ratios for an isolated intersection with fixed flows). In general it is necessary to implement algorithms to perform a local search along a feasible direction, i.e. a direction moving towards a local optimum. The solution reached will be the global optimum only if the objective function is convex. However it is impossible to demonstrate convexity of the objective function for most network design problems. The general scheme of a feasible direction nonlinear optimization algorithm is presented in Fig. 9.4.5.

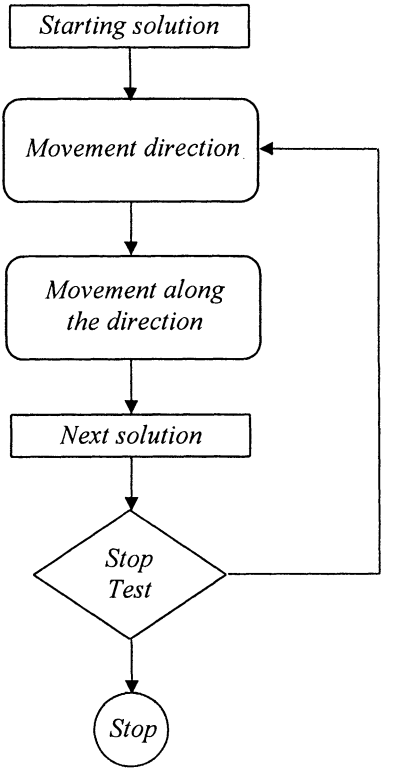


Fig. 9.4.5. General scheme of feasible direction algorithms for continuous SDP.

Different algorithms can be specified according to movement directions. If the movement direction can be shown to be an ascent or descent direction (e.g. the gradient or its opposite) the algorithm is *exact*, otherwise it is *heuristic*. The

movement along the direction can be performed by a linear search or by a fixed or variable step length according to the computational difficulty. Two examples of gradient algorithm applied to continuous network design problems (signal setting and transit line frequency optimization) are reported in the following.

An algorithm for optimal signal setting. A first example of continuous SDP is the global optimization of traffic signal settings for urban networks. Under the assumption that each intersection (node n) has only two phases (a and b) the control variables are the effective green time - cycle length ratios, one for each intersection:

$$y_n^{PER} = G_n^a/T_{c\,n} = 1 - (G_n^b/T_{c\,n}) \qquad \forall \text{ intersection } n$$

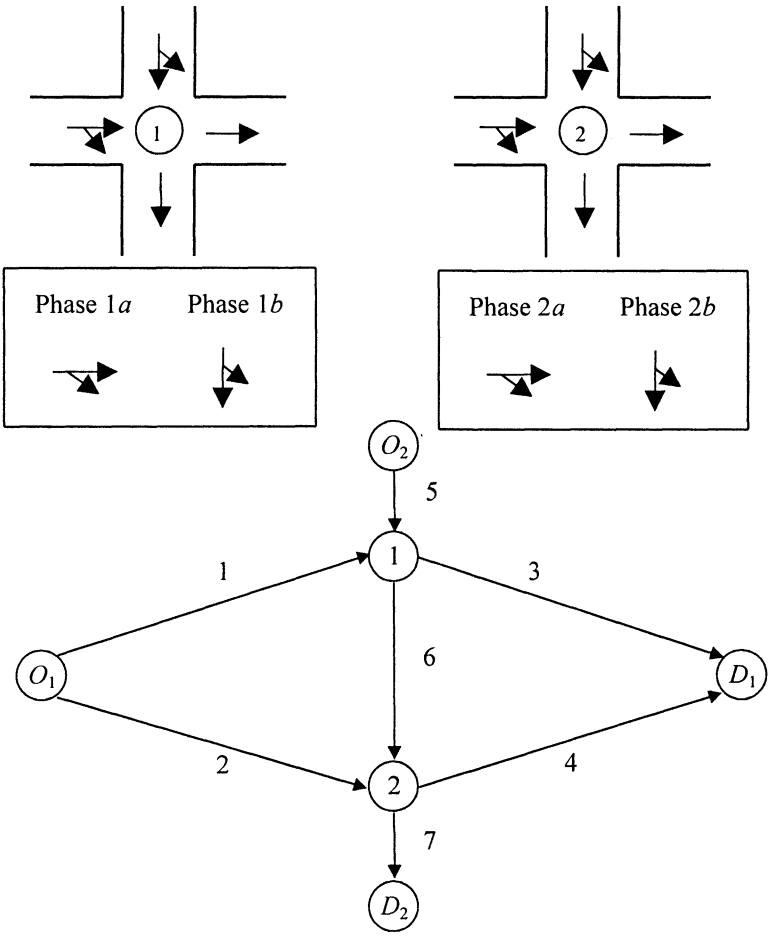
where:

G_n^a is the effective green time for phase a at intersection n;
G_n^b is the effective green time for phase b at intersection n;
$T_{c\,n}$ is the cycle length for the intersection n.

For such a problem the number of variables is equal to the number of signalized intersections since there is only one decisional variable for each of them. The social objective function to be minimized can be the total user cost on the network given by equation (9.3.2).

To solve this problem a projected gradient algorithm with numerical calculation of derivatives and variable step length can be used; the algorithm follows the general framework reported in Fig. 9.4.5 and computes the descent direction as the opposite of the numerical gradient. The descent direction is projected (i.e. some components set to zero) if next solution violates a constraint. In order to find the step length, the descent direction can be normalized by dividing its components by the maximum absolute value.

The algorithm proceeds with a fixed step length each time the new value of the objective function is improved and the constraints on the variables are satisfied. The step length will be reduced each time the objective function value worsens at an iteration and the algorithm ends when the step length is less than a fixed value. A numerical example of the optimal signal setting problem was performed for the small network with two controlled intersections described in Fig. 9.4.6a. The assignment model used in the example is Stochastic User Equilibrium with Multinomial Logit path choice and an MSA algorithm to compute the equilibrium flows (see Chapter 7). The main variables generated by the projected gradient algorithm are presented in Fig. 9.4.6b.



Link	Running time	Variable	Saturation flow
1	10	$y_1^{PER} = y_{1a}$	1000 vph
2	8		
3	13	$y_2^{PER} = y_{2a}$	1000 vph
4	10		
5	10	$y_1^{PER} = 1 - y_{1b}$	1000 vph
6	10		
7	4	$y_2^{PER} = 1 - y_{2b}$	1000 vph

Demand
O_1 - D_1 = 700 vph
O_2 - D_2 = 900 vph

Delay function	$tw\ (f_l, y_n^{PER}) = 3 + 5\ [f_l/(y_n^{PER}\ s_l)]^4$
Objective function	$w(y) = [c(y, f^*)]^T f^*$

Fig. 9.4.6a Projected gradient algorithm for the optimal signal setting problem: test network.

Iteration	1	2	3	4	5	6	7	8	9
$y_{1\,it}$	0.500	0.400	0.300	0.300	0.300	0.275	0.275	0.275	0.269
$y_{2\,it}$	0.500	0.401	0.306	0.306	0.306	0.285	0.285	0.285	0.287
f^*_1, f^*_3	292	309	327	327	327	330	330	330	326
f^*_2, f^*_4	408	391	373	373	373	370	370	370	374
w ($\mathbf{y}_{it}$)	138,719	91,315	74,771	74,771	74,771	74,011	74,011	74,011	74,002
Step size	0.100	0.100	0.100	0.050	0.025	0.025	0.013	0.006	0.006
$\partial w/\partial y_1$	378,527	94,661	35,000	35,000	35,000	4,000	4,000	4,000	- 1,000
$\partial w/\partial y_2$	373,340	88,501	29,000	29,000	29,000	- 1,000	- 1,000	- 1,000	- 2,000
max $\lvert \partial w/\partial y_n \rvert$	378,527	94,661	35,000	35,000	35,000	4,000	4,000	4,000	2,000
Δy_1	- 0.100	- 0.100	- 0.100	- 0.050	- 0.025	- 0.025	- 0.013	- 0.006	- 0.006
Δy_2	- 0.099	- 0.096	- 0.083	- 0.041	- 0.021	+ 0.006	+ 0.003	+ 0.002	- 0.003
$y_{1\,it+1}$	0.400	0.300	0.200	0.250	0.275	0.250	0.262	0.269	0.263
$y_{2\,it+1}$	0.401	0.306	0.223	0.265	0.285	0.291	0.288	0.287	0.284
w ($\mathbf{y}_{it+1}$)	91,315	74,771	84,864	74,864	74,011	74,150	74,030	74,002	74,053
Step size red.	NO	NO	YES	YES	NO	YES	YES	NO	YES
Stop test	NO	NO	NO	NO	NO	NO	NO	NO	YES

VALID Steps | OPTIMAL Solution

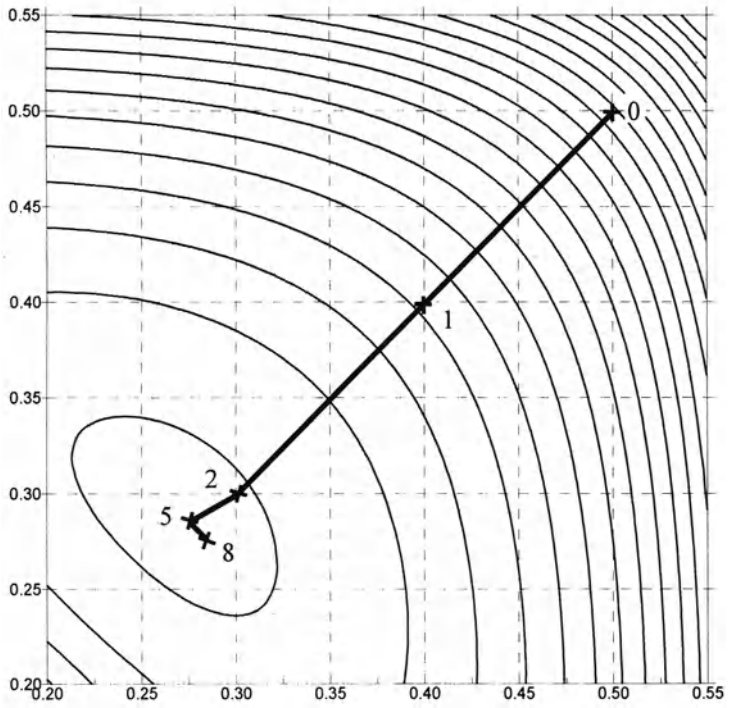
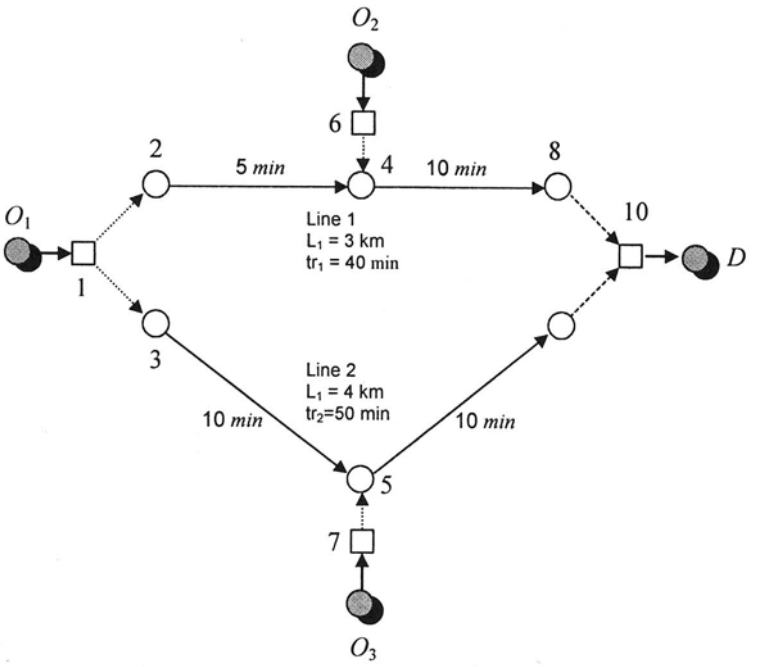


Fig. 9.4.6b - Projected gradient algorithm for the optimal signal setting problem of Fig. 9.4.6a.

An algorithm for optimal transit frequencies. Another example of continuous network design problem is the optimization of transit frequencies.



Demand:
$d_{O1,D}$ = 300 pass/h
$d_{O2,D}$ = 200 pass/h
$d_{O3,D}$ = 100 pass/h
Max bus number:
N_{max} = 10
Constraints:
$1 \le y_1^{PER} \le 10$
$1 \le y_2^{PER} \le 10$
$Int(TT_1 y_1)+Int(TT_2 y_2) \le N_{max}$
Values of time:
Waiting: C_w = 3000 units/h
On board: C_b = 1000 units/h
Kilometric costs:
CE_n = 5000 units/h

Fig. 9.4.7a Projected gradient algorithm for the optimal transit frequency problem: test network.

NUMERICAL RESULTS						
ITERATIONS →	1	2	3	4	5	
$y_{1\,it}$	3	5	7	7	8	
$y_{2\,it}$	3	3,92	4,78	4,78	4,41	
$f_{2\text{-}4}$	150	168	178	178	193	
$f_{4\text{-}8}$	350	368	378	378	393	
$f_{3\text{-}5}$	150	132	122	122	107	
$f_{5\text{-}9}$	250	232	222	222	207	
Operator cost	105.000	153.387	200.506	200.506	208.268	
Users cost	587.700	433.621	360.288	360.288	349.536	
Total costs (w^k)	692.700	587.008	560.794	560.794	557.805	
Bus number	6	8	9	9	10	
Step size	2	2	2	1		
$\partial w/\partial y_1$	-78.500	-21.492	-4.580	-4.580		
$\partial w/\partial y_2$	-36.084	-9.198	1.657	1.657		
Max $\lvert\partial w/\partial y_i\rvert$	78.500	21.492	4.580	4.580		
$\Delta y_{1\,it}$	2	2	2	1		
$\Delta y_{2\,it}$	0,92	0,86	-0,72	-0,36		
$y_{1\,it+1}$	5	7	9	8		
$y_{2\,it+1}$	3,92	4,78	4,05	4,41		
Total costs (w_{it+1})	587.008	560.794	558.624	557.805		
Bus Number	8	9	11	10		
Step size red.	NO	NO	YES	NO		
Stop test	NO	NO	NO	YES		

Fig. 9.4.7b Projected gradient algorithm for the optimal transit frequencies problem of Fig. 9.4.7a.

This problem looks for the optimal frequencies y^{PER}_j for a transit network, with given transit lines. The objective function is the sum of user and operator costs expressed by (9.3.3) and (9.3.4) respectively taking into account only frequency control variables y^{PER}. The constraints included in this model are the assignment constraints (9.3.5), the minimum and maximum frequency constraints and the vehicle budget constraint (9.3.6).

In Fig. 9.4.7 numerical results of an application of the projected gradient algorithm on a test network are shown. The step length can be reduced if the objective function increases, and/or if the budget constraint is violated.

Reference Notes

There is large body of literature on network (supply) design models. For partial overviews, see Oppenheim (1994), Friesz (1985) and Magnanti and Wong (1984) for discrete variables problems. The general formulation proposed in this chapter is original. Fisk (1984) proposes the interpretation of the supply design problem in the context of game theory.

For each application area mentioned there are several articles. See Magnanti and Wong (1984) for networks topology design problems, Marcotte (1983),

Cantarella, Improta and Sforza (1991) and Cascetta, Gallo and Montella (1998, 1999) for the optimal capacity design in signalized urban networks, Davis (1994) for the optimal capacity design in extra-urban road networks, Le Blanc (1988), Nuzzolo and Russo (1997), Montella, Gallo and Amirante (1998) for the design of transit lines and frequencies, Cascetta and Rostirolla (1989) for the design of optimal social fares of railway services, Hearn (1997) for the design of efficiency prices with rigid demand for road network, and Ferrari (1995) for the design of efficiency prices of urban parking systems with elastic demand with respect to modal choice.

The works of Cantarella, Viola and Vitetta (1994) and Montella and Gallo (1998) are examples of applications of design techniques for complex sets of projects (topology of the road network, parking, public transport) in urban areas.
Most of the referenced papers propose heuristic algorithms to solve the related problem. A general review of some algorithms for discrete variable problems is given in Magnanti and Wong (1984). Other examples of algorithms for the discrete network design problem can be found in the papers of Billheimer and Gray (1973), Poorzahedy and Turnquist (1982) and Boyce and Janson (1980). The papers of Le Blanc (1975), Foulds (1981) and Chen and Alfa (1991) are examples of exact branch and bound algorithms.

Algorithms for *continuous network* design problems can be found in Abdulaal and Le Blanc (1979), Marcotte (1983) and Le Blanc and Boyce (1986) for the topological design, and Sheffi and Powell (1983), Heydecker and Khoo (1990), Yang and Yagar (1995) and Cascetta, Gallo and Montella (1998, 1999) for the signal settings design problem.

Notes

(1) In the literature, supply design problems and the relative models are often denominated network design problems (NDP). This definition, as will be seen, is appropriate only for a wide subset of supply design problems, which refer to the definition of network elements.

(2) The model described corresponds to a monopoly market. In reality, the situation is often more complex. For example, in the transportation market there might be many operators (e.g. air service, railway and road managers) each with his/her own objectives and constraints and the ability to forecast the demand reactions to his/her own actions and to those of the competitors. The supply design models available are not yet capable of simulating this type of market defined as an oligopoly.

(3) The two objective functions (9.2.4) and (9.2.5) coincide only in the case of deterministic path choice model.

(4) The economic interpretation of this objective function, differing from the total generalized cost, is that the monetary cost can be considered a transfer from users to system operators which, in principle, can return it to the users in another form (see section 10.5.1 on Benefit-Cost analysis).

10 TRANSPORTATION SYSTEMS ENGINEERING FOR PLANNING AND EVALUATION

10.1. Introduction

Transportation systems engineering can be defined as a discipline aimed at the functional design of physical and/or organizational actions on transportation systems. Each set of coordinated, internally consistent actions is referred to as a project or plan. The transportation system engineer must also evaluate the main potential effects of the project(s) to test their technical suitability and to support intermediate and final decision-makers.

The scope of transportation system projects might be very different, as are the various points of view from which their consequences can be evaluated. Projects might relate to transportation facilities, control systems, services and fares. Each can be designed and evaluated from the perspective of the community served by the transportation system under analysis, or from the perspective of the service and/or facility operators. Design and decision-making are two interdependent activities. Decision-making for transportation systems is usually more complex than for many systems analyzed and designed by other sectors of engineering. This is especially true when the decision-maker must, either directly or indirectly, consider the effects of proposed actions on the collectivity. Projects concerning decisions and/or points of view typical of the operator, such as the organization of freight distribution or the design of a traffic light control system, usually undergo a relatively simpler and more straightforward decision process. Often, also company projects, such as the reorganization of transit lines, lead to impacts "external" to the company which may influence the final decisions. For this reason, this chapter will refer mainly to complex projects with a wide range of impacts.

The following sections will first describe the role of transportation system engineering in the context of the widest decision-making (planning) process (section 10.2) as well as some fields of application (section 10.3). Then reference will be made to the activities relating to project evaluation and, in particular, to the phases of identification of the relevant impacts and their quantification using the models and methodologies described in the previous chapters (section 10.4). Finally, in section 10.5, some elements of the techniques most frequently used for the

E. Cascetta, *Transportation Systems Engineering: Theory and Methods*

comparison of alternative projects (Cost-Benefit and Multi-Criteria analyses) will be discussed.

10.2. Transportation systems engineering and the decision-making process

In Chapter 1, it was stated that changes in transportation systems may affect a community and its members in several ways. Building a new facility, for example, may not only change service performances for network users, but also produce economic, financial, social and environmental effects for many groups or individuals who are not system users. These non-users may be single individuals as well as entrepreneurs, landowners, operators and institutions responsible for the transportation system and the area in which it operates.

Project decisions can be made in many different ways. The "rational" approach to decision-making is based on the evaluation of the various effects of the different possible projects on the different parties involved. This approach, which is commonly adopted in the case of "private" decisions, is even more necessary when the decisions are made on behalf of a community. The natural dynamics of society, economic cycles, changes in individual's and decision-makers' attitudes, the occurrence of particular events, the availability of resources are such that decisions and their implementation evolve over time. This has resulted in changes over the years in the very concept of planning. Planning is no longer seen as the draft of a single plan, or as a "closed" activity defining projects to be implemented over a sufficiently long period of time. A *planning process* is a sequence of decisions (plans or projects) taken at different, not necessarily predefined, moments in time accounting for the effects of previous decisions. In this framework the role of quantitative methods for the definition and the evaluation of alternative projects is even more relevant as they ensure a sort of "dynamic rationality" to the whole process.

The theoretical analyses that have produced a "planning theory" as a theory of collective decision-making are beyond the scope of this book. However the identification of the role and the limits of transportation systems analysis and engineering within the wider decision process is extremely relevant. To this end it is useful to consider schematically the different macro-activities of the decision process, see Fig. 10.1.1. The right side of the figure shows the decision process, while the left side shows the phases of analysis and modeling functional to these activities.

In the phase of *objectives and constraints identification*, the objectives of the decision-maker (or decision-makers) and the relevant constraints for the project are defined. Objectives and constraints may be explicit or, at least partly, implicit. They depend on the perspective of the decision-maker and, in one way or another, define the type of actions that can be included in the project (e.g. new facilities over the long term or the reorganization of the existing facilities in the short term).

In Chapter 9 it was shown that modifications to the transportation system can be designed from different points of view. Among the typical objectives of an operator there should be the maximization of net profit. Constraints might include the existing regulations, the available budget, service and/or fare obligations, the technical limits on the production capacity of the factors employed, etc. In the case of public decision-makers, the project objectives are many, often not clearly defined and conflicting with each other, as are the interests of a "complex" society. A public decision-maker may be interested in increasing safety, reducing the generalized transportation cost borne by the users, increasing equity in the distribution of transport benefits, improving accessibility to economic and social activities, fostering new territorial developments, protecting environmental values, and reducing the public deficit. Objectives and constraints, explicit or implicit, synthesize the values and attitudes of the firm or of society. The increasing relevance of energy consumption and environment preservation in recent decades are clear examples of this point.

Both objectives and constraints influence the successive phases of the process and in particular the analysis of the present situation and the actions that can be included in alternative projects. From the modeling perspective, these factors impact *the definition of the system of analysis*, i.e. the identification of the elements and their relationships included in the representation of the system in order to evaluate correctly the effects of planned actions.

In the phase of *analysis of the present situation*, data on the transportation and activity systems are collected. Data are used for the analysis of the present state of the system and for the identification of its main inadequacies or "critical points" with respect to the objectives and the constraints of the project. In a problem-solving approach, the critical aspects should be corrected or alleviated by the planned actions. This phase is also linked to the *building of a mathematical model of the present system,* since it provides the input data for the models (supply, demand, land-use). Furthermore it usually receives from the models estimates of some system performance indicators (e.g. flows, saturation levels, generalized transport costs by O-D pair) impossible or too costly to measure directly.

The next step is *the formulation of system projects (or plans),* i.e. sets of complementary and/or integrative actions which are internally consistent and technically feasible[1]. The strict interdependence among the elements of a transportation system generally requires a project be designed taking into account the various components that may be significantly influenced by it. A new subway line, for example, requires a reorganization of the surface transit lines to increase the catchment area of the stations (complementary action). Restricting the access of cars to parts of an urban area requires the design of appropriate parking areas, transit lines, pricing policies and so on (integrative actions). Systems design is usually limited to the definition of the functional characteristics of the elements composing the system; their physical design, if necessary, pertains to other branches of engineering.

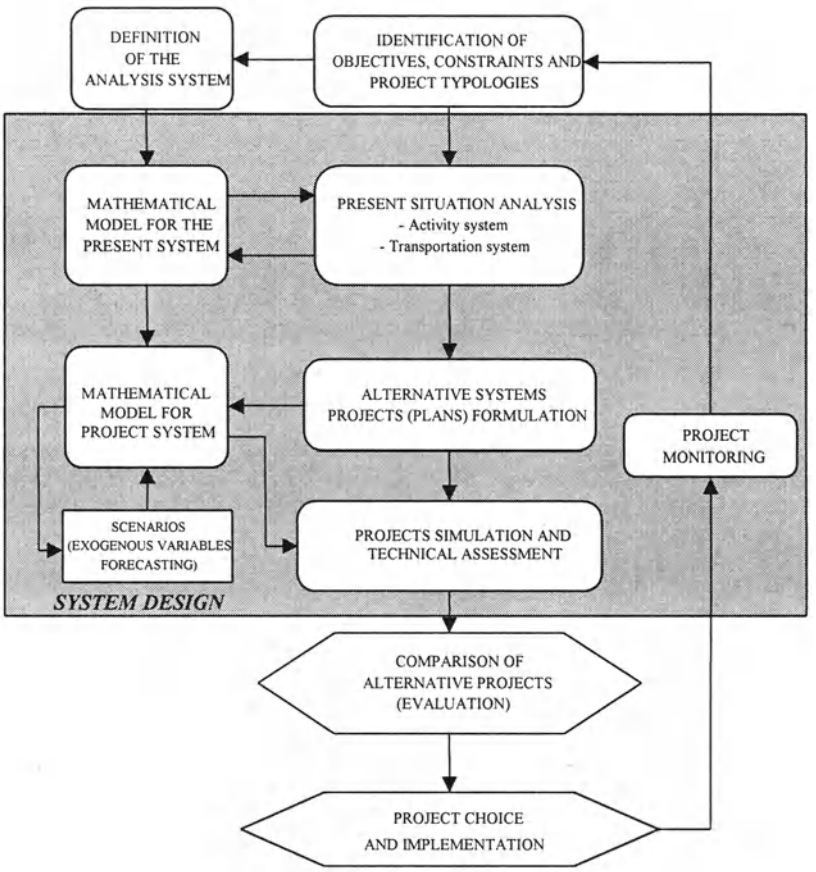


Fig. 10.1.1 –Transportation systems design and the planning process.

In general, several alternative projects can be proposed in response to predefined objectives. One alternative is the non-intervention (do-nothing) option, i.e. the possibility to keep the system in its present state or, more realistically, to follow the decisions already taken. For more complex projects requiring several actions, which cannot be implemented simultaneously, alternative time sequences can be generated and each sequence can be considered as an alternative project. In fact, the effects, and the "convenience", of a project may be significantly influenced by the specific sequence of actions implemented.

Assessment and evaluation of alternative projects require *the simulation of the relevant effects (impacts)* of their realization. Most of the impacts can be simulated quantitatively using the mathematical models described in previous chapters. Supply

design models described in Chapter 9 can also be used as tools to generate alternative configurations to be compared and evaluated, especially for simpler design cases. If the evaluation of a project requires the simulation of its main impacts on a sufficiently long time horizon, assumptions on the "future" structure of the activity system, or rather on the values of the variables exogenous with respect to the model adopted, are needed. A set of consistent assumptions on the activity system is usually known as a *socio-economic scenario*. The evolution of exogenous variables over long time periods depends on complex phenomena related to the demographic, social and economic evolution of the area and on the related external environment. It is very difficult, if possible at all, to forecast these phenomena with sufficient precision. Thus the usual practice is to consider different scenarios to estimate the range of variation of the simulated effects and to check the robustness of the alternative projects with respect to the possible future scenarios.

The *technical assessment of the projects* concludes the system design phase. This activity verifies that the elements of the supply system are "functional" within their ranges of economic validity and technical feasibility (e.g. that the forecasted users' flows are not too low or too high with respect to their technical capacity). Moreover the technical feasibility of supply performances assumed and their consistency with the simulated system state are checked. Technical assessment is performed on the basis of simulated impacts of the projects. Simulations can (and often do) have feedbacks on the formulation of projects as it is usually the case with the design of engineering systems[2].

The activities related to the analysis of the present situation, the formulation of alternative projects, the simulation of relevant effects and the technical assessment can be collectively defined as the *system design phase*.

The effects of alternative projects can be further processed to facilitate their comparison. There are many techniques for the analysis and *comparison of alternative projects* with different levels of aggregation, such as Cost-Benefit and Multi-Criteria analyses. However, it should be stressed that these techniques cannot and should not replace the actual decision-making process, which is based on compromises among conflicting interests and objectives; rather they should be considered as tools to support actual decision-making.

After a project is implemented, one can compare forecasted and actual effects, observe the occurrence of unexpected developments and new problems, and evaluate social consent and/or dissent. These may modify some elements of the project or alter its future development. The *monitoring*[3] of a project is the systematic checking of the main "state variables" of the transportation system and use of these checks for the identification of new problems and the a posteriori evaluation of project impacts. In practice, monitoring transportation systems and projects is often neglected or carried out non-systematically, although it should play a much more important role in the planning process.

The complexity of the decision-making processes for transportation systems is clear from what has been said so far. The analyst has a technical role in the phases of analysis, design and simulation of the interventions. It should also be recognized that

in general the transportation systems engineer does not have the technical skills required for all the tasks involved. Interactions with specialists from other disciplines such as the other branches of engineering, economics, urban and regional planning as well as social sciences are needed, particularly if the projects imply significant effects on external systems. On the other hand understanding the "functioning" of transportation systems and therefore their design and quantitative simulation are the core of the disciplinary competence of transportation systems engineers.

10.3. Some areas of application

The decision-making process described follows the "rational" model, a model that is often considered to be a gross simplification of public decision-making processes in the real world. In spite of this criticism it should be seen as a reference paradigm which, with necessary adaptations, can theoretically be applied to very different problems and decisional contexts. Some examples of applications for transportation system engineering will be discussed below together with their implications on the mathematical models.

Strategic transportation planning

Strategic or investment planning involves decisions on long-term (10-20 years) capital investment programs for the realization of new infrastructures (e.g. roads, railways, ports) and/or the acquisition of vehicles and technologies (e.g. rolling stock and control systems). In this case, projects usually include transportation services, pricing policies and, in some cases, travel demand management policies (e.g. access or parking restrictions). Public projects result in urban, regional, national or transnational transportation plans, depending on the extent of the area, while company-oriented projects result in a strategic company development (or business) plans.

For strategic plans, the whole transportation system is usually considered to be the study system because substantial changes, even for a single mode, may influence the structure of the whole system. Returning to the example of an urban transportation plan for a new subway line, the design elements will also include the surface transit lines, the parking policy, the fares policy, etc. The evaluation of its effects cannot be limited to the public transportation system since it is very likely that the modal split of demand will change with significant effects on road congestion, parking availability and so on. The temporal horizon for this level of design requires the forecasting of scenarios for the activity system. Furthermore, the inverse interaction between the transportation system designed and the activity system should be considered as well. Continuing with the same example, it is reasonable to expect that the construction of a new subway may change, at least to some extent, the land use pattern and therefore transportation demand. This wider view of the design system usually is associated with a less detailed representation. In fact, it is irrelevant to simulate extremely detailed effects such as turning maneuvers

at intersections or flows on minor roads since they are not significant for the evaluation of the project under study.

Feasibility studies of transportation projects

These applications fall within the context of "programming by projects" based on the formulation of a reference scheme identifying the relevant connections and the subsequent evaluation of the projects related to individual connections in order to assess their technical feasibility, economic convenience, priority level and mode of realization.

Technical and economic feasibility studies of transportation facilities usually require the formulation of alternative system projects defining the performances and functional characteristics of the connection (such as layout, connections, capacity, service performances, type and characteristics of vehicles and technologies, prices). Alternative projects, including the do-nothing or reference solution, are then evaluated from the functional, economic and financial point of view in the context of different transportation and activity systems scenarios. Also in this case the temporal horizon is usually long-term; the scale varies from urban to regional or national according to the kind of project to be assessed. The system under analysis can be analyzed and modeled with spatial and functional levels of detail differing according to the intensity of the interrelations with the connection under study. For example, a denser zoning system can be adopted around the alignment of a new railway. Whatever the case, the system must be simulated with reference to the travel demand and the supply of all transportation modes.

There are several examples of feasibility studies, both urban and extra-urban. Some of the studies are aimed at assessing the convenience of private capital investments in facilities and/or transportation services (project financing). In this case forecasts of travel demand, users' flows and revenues are of special interest, as well as the "external" conditions under which expected demand and profits can be obtained.

Tactical planning

Short/medium term tactical planning is concerned with decisions on projects requiring limited resources, usually assuming minor or no changes in the infrastructures. Urban traffic plans or public transport plans are examples of tactical plans under the public point of view. The design of scheduling and/or pricing policies for air or rail services are examples of tactical plans from the operators' point of view.

In this context, the evaluation of the technical and functional effects of the project, as well as the financial analysis in terms of operating costs and traffic revenues are of primary interest. These analyses might be accompanied by an economic evaluation, though often simplified. For these applications, the socio-economic scenario is usually assumed to be given. In practice, it is also assumed that the level and spatial distribution of travel demand are unaffected by the projects, while variations in modal split and assignment to the networks, representing the

transportation services involved, are explicitly simulated. In some cases a single transport mode is examined in the context of the overall system, therefore taking into consideration the effects of "modal competition" only through the level of demand of the mode considered (elasticity analysis), without an explicit representation of the supply of competing modes.

Operations management programs

Short-term operations management programs generally define particular aspects of individual mode operations, optimizing the use of the available resources usually from a company point of view. Traffic-signal control plans, design of transit timetables, and organization of factors necessary for producing transportation services (e.g. the assignment of vehicles to lines and travel staff to work shifts) are examples of operations management programs.

In this case the study system is usually limited to a single mode assuming that the modal demand is fixed. For example, only the road sub-system (network and demand) is considered in designing a traffic-signal control scheme. If necessary, network and assignment models described in previous chapters can be integrated with detailed micro-simulation models. Furthermore, the design phase can be carried out with the support of supply design models similar to those described in Chapter 9.

10.4. Evaluation of transportation system projects

Project evaluation can be defined as the assessment and comparison of the available alternatives on the basis of their effects with respect to the objectives and the constraints of the decision-maker. As stated previously, evaluation is (or should be) a technical activity carried out by the analyst interacting with the parties involved with the ultimate aim of supporting decisions. On the other hand, choice is essentially an activity of synthesis and mediation among conflicting interests. Choices are made by one or more decision-makers, both formal and informal, interacting in a complex institutional context. For the sake of description, it is useful to maintain the conceptual division between evaluation and decision-making, though in reality there are inter-relationships, often close, between the two phases. Better awareness of the consequences can, for example, modify the objectives or, more generally, the attitude of the decision-makers.

As an activity supporting decisions, evaluation depends on the decision-maker's perspective. A classic example is the difference between the financial and economic analysis of a project. *Financial analysis* attempts to maximize profit under constraints such as regulations, service obligations, concessions, etc. In this case "benefits" and "costs" can be expressed in monetary terms; the former come from the revenues from service sales and subsidies, if any, the latter from the financial costs of service production such as construction, maintenance and running costs, tolls, taxes, etc. *Economic analysis* is traditionally associated with a public decision-maker[4]. Alternative projects are evaluated taking into account positive and negative impacts (benefits and costs) with respect to the objectives of the collectivity, or rather of the different groups homogenous in terms of their socio-economic

characteristics and of the impact received. Some transportation system users may in fact benefit from a particular project (reduction of travel times and costs, increased accessibility, etc.) while others may have lesser advantages or even disadvantages (increased travel times and costs, etc.) This might occur, for example, in an urban area as a result of the migration of congestion from one zone to another due to traffic signal control strategies, reserved lanes for public transport, limited access traffic zones, etc. The contrast is even more evident if the benefits to system users are compared with the costs borne by some non-users, for example the increase in noise and air pollution for residents in zones close to a new motorway or a new airport.

Many techniques have been proposed for the evaluation of transportation system projects. In general, the evaluation can be decomposed into three logically successive phases:

a) identification of the effects, or impacts, relevant to the formal and informal actors in the decision-making process and related to alternative system projects;
b) identification of the quantitative and qualitative variables (impact indicators) representing the impacts and estimation of their variations included by each project;
c) comparison of alternative projects on the basis of their respective impacts.

Quantitative evaluation techniques of public transportation projects have been the object of many theoretical studies and practical applications over the decades. From the end of the '50s, when they were applied to motorway projects, this discipline rapidly evolved both from the theoretical and practical point of view. In recent years the aim and scope of project evaluation have grown following some major changes in the transportation arena. These include changes in values and participation of different interest groups, deregulation of some sectors of the transportation market, the involvement of private capital in financing infrastructure construction and/or service operations. The systematic analysis of the results achieved in this field is well beyond the scope of this book. The following sections will consider the three phases listed above only in order to identify the role of quantitative methods in the overall activity of projects evaluation.

10.4.1. Identification of relevant impacts

Impacts of a transportation system project can be defined as the consequences of the project relevant for some of the actors involved (i.e. groups of individuals homogeneous with respect to the issue under consideration). Thus the definition of the relevant impacts is the main indicator of the approach followed and the breadth of the evaluation activity. The spectrum of the effects considered has widened with the passing of time in concert with improvements in models and computing power and with the expansion and classification of the different and often contrasting objectives and goals of actors and decision-makers. The same figure of the "decision-maker" has been defined and differentiated in step with variations in the financing and management of transportation systems. The generic and undifferentiated "public operator", typically imagined as a state or local agency

reconciling general public and operators objectives, has led to a separation of the different roles, in particular those of public institutions, main interest groups and transportation services operators independently of their "public" or "private" nature.

The first generation of quantitative evaluation exercises initiated in relation to investment in motorways and later extended to transport system project in the broader sense. They took into account only the monetary and monetarily quantifiable effects (benefits and costs) for the *users of the planned facilities* and for *building and operating* these facilities and services. The former included the variations[5] in level-of-service attributes such as travel time, monetary cost of tolls and vehicle operation; variations in the expected number of accidents were sometimes included in the evaluation. The monetary cost for *service and/or infrastructures* operators included the construction costs, investment costs in vehicles and technologies, variations of maintenance and running costs as well as the variations in revenues from service sales. The effects for the operator sometimes included variations in transfers with other higher-level public authorities (e.g., reimbursements for service obligations, duties and taxes on gasoline and on premises, etc.)

With better understanding and modeling of the mechanisms underlying transportation systems, the range of effects considered for *the users of the transportation system* gradually increased. The impacts are considered for all users, both present and project-induced, calculating the variations in generalized costs, perceived and not perceived, for the different transportation modes. Often the impacts are differentiated for the different classes of users (or market segments), i.e. for groups of users homogeneous in terms of trip purpose, socio-economic characteristics and level-of-service attributes. As for the effects on operators, construction, maintenance and running costs calculated on the basis of market prices have been decomposed gradually by the resources employed (manpower, materials, capital) since market prices do not always reflect the actual social "value" of the resources.

A further widening of the prospective as well as of the spectrum of the effects considered in the evaluation of a transportation project relates to the "*external effects*" of the project. These effects relate to those members of the society not directly involved in the use of the transportation system. Some examples of the impacts on non-users will be given below subdividing the external effects into economic, territorial, social and environmental. It should be noted that the classification of some impacts can be somewhat arbitrary and there is no generalized consensus among the analysts.

Economic impacts can be defined as changes in the state of the economic system brought about by the project. Variations of residential and commercial property values and in economic production following variations of accessibility; variations of the economic impacts of accident directly and indirectly connected to the project can be listed in this group. Economic externalities are directly measurable in monetary units, or at least can easily be translated into such units.

Territorial impacts are related to land use and its quality. Examples of territorial impacts are variations in the use of property (e.g. from residential to commercial) or more generally the relocation of housing and economic activities brought about by accessibility differentials. Within this class are changes in the geographical structure of a region or in the urban quality of certain neighborhoods.

Social impacts can be defined as impacts on social values and variations in the relationships among people and social institutions such as the family, local communities, education, government bodies, etc., brought about by the project. In this case too there are effects of different types: social effects of accidents, variations in accessibility to social activities (schools, public offices, parks, etc.), changes in cohesion and stability of local communities, impacts on historic and cultural sites. Variations in equity, e.g. changes in the distribution of travel related opportunities with respect to space (zone) and socio-economic status (income class or age) can also be considered as social impacts.

Finally, *environmental impacts* can be defined as the effects of the project on the physical environment. These can be classified as effects on the ecosystem, on noise and air pollution, and on visual perception. Transportation system projects, especially in the case of new infrastructures in rural areas, can alter the ecological equilibrium of vegetation and animal populations. Furthermore, any transportation system generates noise and air pollution. The project may significantly change the intensity and the distribution of pollution. Visual impacts, lastly, are direct effects of infrastructures and vehicles dependent on their "visibility" and on their "contrast" with the surrounding background.

Figure 10.4.1 synthesizes some of the effects of a transportation system project for the different groups involved. It is obvious that not all the impacts listed are relevant to the evaluation of all projects. In certain cases, some effects might be absent or their variations can be considered negligible; in other cases some impacts may be present but irrelevant to the particular point of view of the analysis.

10.4.2. Identification and estimation of impact indicators

The effects of a transportation system project are usually represented by a set of variables known as *impact indicators* or *measures of effectiveness* (MOE). Since, in general, there are several elemental impacts, and it is impractical to handle all the related variables, it is common practice to use for further analyses a reduced number of performance indicators obtained as aggregate variables, or "intermediate constructs".

Some impact indicators are quantitative variables such as travel time or CO tons of gas emissions; others are "structurally" qualitative and can, at most, be expressed by descriptive variables (adverbs such as 'little', 'much', etc.) or on an arbitrary scale (such as from A to F).

* *Users (by class)*
- Differences of net utility (surplus) perceived by the users
- Differences of costs non-perceived by the users

* *Administrations and operators (for each subject involved)*
- Differences in resources (manpower, materials, capital) and in costs needed for building transport infrastructures, vehicles and control systems (investments)
- Differences in resources and costs for maintenance of the infrastructures and technologies
- Differences in resources and costs for the operation of transportation services
- Expropriation and re-allocation costs
- Differences in traffic revenues
- Variations in taxes paid by users (fuels, etc.) and non-users (property, etc.)
- Differences in transfers between administrations

* *Non-users of the transportation system (for each homogenous group)*

Economic Impacts
- Differences in the production of different economic sectors
- Differences in the economic impacts of accidents
- Differences in property values

Territorial Impacts
- Differences in the location of households and economic activities
- Differences in urban structure and "quality"
- Impacts on the preservation of historic and cultural sites

Social Impacts
- Difference in accessibility to social activities (school, social and religious centers, recreational activities, etc.)
- Modifications in the structure and cohesion of local communities
- Variations in the social effects of accidents
- Variations in the distribution of users'surplus by zone and and socio-economic group (impact on equity)
- Variations in visual and aesthetic impacts

Environmental Impacts
- Changes in the ecosystem
- Variations in noise and air pollution

Fig. 10.4.1 – Classification of impacts for the evaluation of transportation system projects.

The effects of a project are usually evaluated in differential terms, i.e. as variations or differences of the variables representing them, between the project (*P*) and non-project (*NP*) states. The latter, sometimes known as "reference solution", is defined as the option to maintain the present state of the system, or to go along with the projects already decided which are not subject to the evaluation.

The *time dimension* is an important factor in estimating impacts. The impacts of a project occur in time following different profiles. For example construction and investment costs are spent in a relatively short period of time while maintenance and operation costs continue throughout the entire life of the project. Furthermore, with the passing of time some effects change in intensity or even in direction: the travel cost perceived by the users may increase during the construction phase of a given facility due to capacity reductions, and other disturbances, while they decrease when

the infrastructure is operating. Conventionally the economic life [6] of the project is decomposed into stationary reference sub-periods, for example, different periods of the year and different time bands of typical days for congested systems, both during construction and representative operation years. As was seen in Chapter 1, impact indicators are typically computed for a subset of simulation or reference periods and then extrapolated to larger time periods. Many impacts can be simulated by using the models described in previous chapters as shown in Fig. 10.4.2. The estimation of these impacts requires the simulation of the system in the project (*P*) and non-project (*NP*) states and the calculation of differences (variations) between the variables measuring quantifiable impacts. As stated in section 10.2, for long-term projects, each simulation will require a set of coherent assumptions on the evolution of the exogenous variables (scenarios).

Resources needed for *construction, maintenance* and *operation* and their relative costs can be estimated analytically from the actual design of facilities and services or, synthetically, using statistical relationships known in the economic literature as *production functions.* The latter estimate the resources needed to build and equip a unit length of typical infrastructure, produce vehicles and technologies with given characteristics, maintain infrastructures and operate a transportation service of a given type. Alternatively, *construction, maintenance* and *operation cost functions* estimate directly related costs in monetary terms.

Traffic revenues can be calculated by multiplying the number of users in the simulation for tolled infrastructures and/or for transportation services for the relative prices.

Several other impacts can be calculated from the models described. For example, the probability of accidents and their consequences, fuel consumption, noise and air pollution can be evaluated through the relevant link impact functions described in section 2.2.5. The ease of access to different services can be measured through *accessibility variables* deriving from destination choice models, see 4.3.2 or in other forms proposed in the literature. In any case generalized costs or level-of-service attributes play a key role in the measurement of accessibility.

The calculation of the effects perceived by the users of the transportation system requires a further elaboration of the concepts and the demand models described in previous chapters and will be covered in the following section.

10.4.3. Computation of impacts perceived by the users

The impacts perceived by the users can be calculated as a variation of net perceived utility (or surplus) associated to travel choices made in the project (*P*) and non-project (*NP*) states or scenarios. The calculation can be carried out following two different approaches depending on the assumptions underlying the demand model used, i.e. whether they are behavioral random utility models or descriptive, non-behavioral models. In the following the two approaches will be analyzed and compared.

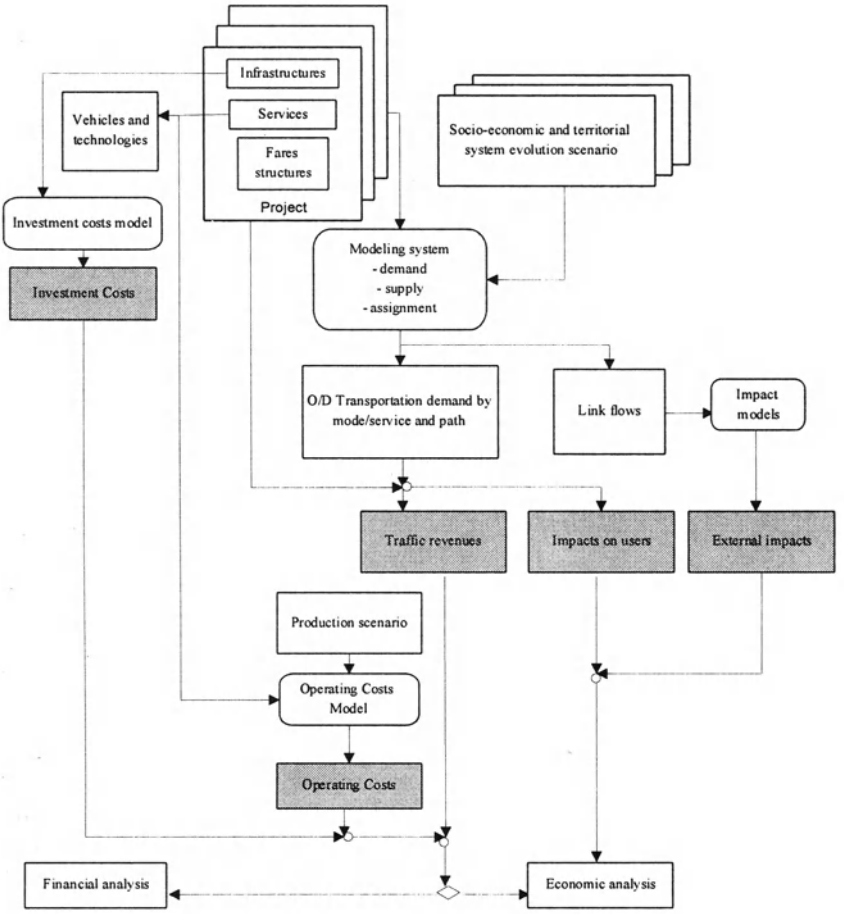


Fig. 10.4.2 – Main components of an impact assessment process.

a) *Random utility demand models*

Random utility demand models are based on explicit assumptions on choice behavior of the generic decision-maker/user *i*. These assumptions can be used to estimate variations in average perceived utilities for relevant choice dimensions. As an example, consider the classic choice sequence on the dimensions "making *x* trips for purpose *s*/ to destination *d*/ with mode *m*/ following path *k*". In this case, the

utility U_P^i perceived by user i in zone o for the sequence that would be chosen in the state P of the system can be expressed as:

$$U_P^i = \sum \beta_k X_{kj(i)}^{iP} + \varepsilon_{j(i)}^i = V_{xodmk}^i(\boldsymbol{X}_{j(i)}^{iP}) + \varepsilon^i{}_{xodmk} \qquad (10.4.1)$$

where $j(i)$ indicates the specific sequence (x o d m k) chosen and the vector of attributes $\boldsymbol{X}_{j(i)}^{iP}$ include level-of-service (times, costs, etc.) and other variables corresponding to $j(i)$ in the project P. Since some attributes $X_{kj(i)}^i$ have positive coefficients (i.e. they represent utilities) while others have negative coefficients (costs), expression (10.4.1) represents the perceived net utility (utility minus cost) or surplus. An elementary specification of the systematic utility V_{xodmk} in the case of shopping trips might be:

$$V_{xodmk}(\boldsymbol{X}) = \beta_1 NOTRIP + \beta_2 SHP_d - \beta_3 t_{odmk} - \beta_4 mc_{odmk} \qquad (10.4.2)$$

where x assumes the values zero and one, *NOTRIP* is a specific variable of the alternative not to make a trip, (x=0), SHP_d is the number of shops in zone d, t_{odmk} and mc_{odmk} are respectively the travel time and monetary cost to go to d with mode m departing from the origin o and following the path k. The linear combination of travel time and monetary cost can be denoted as generalized path cost, $g_{odmk} = \beta_3 t_{odmk} + \beta_4 mc_{odmk}$.

In random utility models, the *perceived utility* (10.4.1) is a random variable and maximizing its value; the alternative chosen and its utility are unknown to the analyst and therefore represented as random variables. Impacts on transportation system users can be expressed by the variation of the *expected value of the surplus* perceived by all the decision-makers of equal characteristics. This value corresponds to the mean of the net utility (surplus) perceived for the alternative chosen, which is that of maximum utility. The mean value of the perceived surplus thus coincides with the mean value of the maximum perceived utility among all the available alternatives, i.e. with the *Expected Maximum Perceived Utility* (EMPU) variable.

Since the specification of the models, the attributes considered and the coefficients β_k in systematic utilities (10.4.1) usually depend on the trip purpose and the socio-economic characteristics of the decision-maker, the EMPU variable s is calculated separately for the generic user of the class i [7] in zone o:

$$s_P(o,i) = E\left[\max_{xdmk} U_P^i(xdmk)\right] \qquad (10.4.3)$$

As shown in Chapter 3, if residuals ε_{xodmk} are i.i.d. Gumbel variables of parameter $\theta = 1$, the EMPU (10.4.3) can be expressed in closed form as a logsum variable:

$$s_P(o,i) = ln \sum_{xdmk} exp\left[V_{xdmk}^i(\boldsymbol{X}_i^P)\right] \qquad (10.4.4)$$

Similar closed form expressions can be obtained for other models belonging to the Logit family. The total net utility of the users of class i in zone o in the project state, $S_P(o,i)$, can be estimated as:

$$S_P(o,i) = N_o^P(i)\, s_P(o,i) \tag{10.4.5}$$

where $N^P{}_o(i)$ is the number of users of class i in zone o. Notice that $N^P{}_o(i)$ includes both actual and potential, i.e. those choosing not to travel in the NP state, users. The variation of perceived surplus for the users of class i in zone o can be expressed as:

$$DS_P(o,i) = S_P(o,i) - S_{NP}(o,i) \tag{10.4.6}$$

where $S_{NP}(o,i)$ is the total perceived surplus in the non-project state.

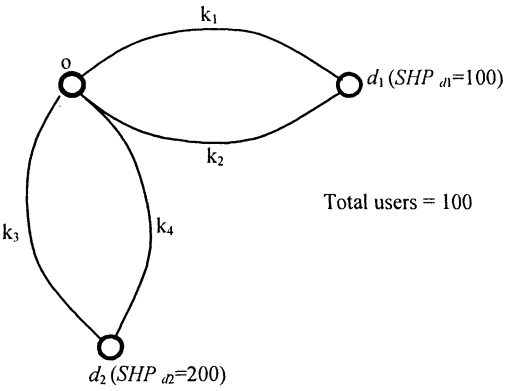
The monotonicity of the EMPU variable, discussed in section 3.5, ensures that the net utility variation may be positive or negative according to whether the systematic utility of each alternative increases or decreases passing from the non-project to the project state. Reductions in cost attributes and/or increases in utility attributes, will lead to increases in the total surplus and vice-versa. For the same property the total surplus will increase if the number of available alternatives increases; this may be the case if the project includes new transportation modes or services. Figure 10.4.3 exemplifies the calculation of perceived surplus for a Logit choice model over the sequence (*xdmk*) for shopping trips.

Users of different classes, in different traffic zones, can be aggregated in various ways, assuming interpersonal summability of utilities. The perceived impact of project P for the aggregate of all users is therefore given by:

$$DS_P = \sum_i \sum_o DS_P(o,i) \tag{10.4.7}$$

It should be noted, however, that in many applications perceived surplus variations should be analyzed for disaggregate user groups in order to highlight the distribution of project benefits among the different groups or zones in the study area.

Average perceived surplus $s_P(o,i)$ and total utility variations calculated by equations (10.4.6) and (10.4.7) are expressed in dimensionless measurement units, sometimes denominated *util*. In order to compare them with other effects of the project P, these values can be expressed in monetary units by dividing them by the coefficient of the monetary cost coefficient β_c, with dimension monetary units^{-1}.



$$V_{xdmk} = \beta_1 NOTRIP + \beta_2 SHP_d + \beta_3 T_{odk}$$

$$S(o) = 100 \ln\left[\exp(\beta_1 NOTRIP) + \exp(\beta_2 SHP_{d_1} + \beta_3 T_{od_1k_1}) + \right.$$
$$+ \exp(\beta_2 SHP_{d_1} + \beta_3 T_{od_1k_2}) +$$
$$\left. + \exp(\beta_2 SHP_{d_2} + \beta_3 T_{od_2k_3}) + \exp(\beta_2 SHP_{d_2} + \beta_3 T_{od_2k_4})\right]$$

	Non- project		Project	
β_1=1	T^{NP}_{od1k1}= 6	g^{NP}_{od1k1} =1,2	T^{P}_{od1k1} = 5	g^{P}_{od1k1} = 1,0
β_2=0.015	T^{NP}_{od1k2}= 7	g^{NP}_{od1k2} =1,4	T^{P}_{od1k2} = 5	g^{P}_{od1k2} =1,0
β_3=-0.2	T^{NP}_{od2k3}= 10	g^{NP}_{od2k3}=2,0	T^{P}_{od2k3} = 6	g^{P}_{od2k3} =1,2
	T^{NP}_{od2k4} = 10	g^{NP}_{od2k4} =2,0	T^{P}_{od2k4} = 7	g^{P}_{od2k4}=1,4

$SNP(o) = 236.2$ $SP(o) = 283.4$ $\Delta S(o) = 47.2$

Fig. 10.4.3 – Calculation of surplus variation with a behavioral model.

If perceived utility variations from the project do not influence trip frequencies, and therefore the demand level for each class of users remains constant (assumption of rigid demand level), the surplus of non-travelers does not change and the total perceived surplus can be expressed as:

$$S_P(o,i) = d_{o.}(i) E\left[\max_{dmk} U^i(dmk / os)\right] = d_{o.}(i) \cdot s_P(o,i) \qquad (10.4.8)$$

where $d_o(i)$ is the number of trips with origin o undertaken by users of class i in the reference period. Similar simplified expressions can be derived for the cases of rigid origin-destination demand flows.

b) Descriptive demand models

A different methodology is adopted to evaluate the impacts for the users in the case of descriptive demand models. In this case the model can be interpreted as a "demand function" relating the number of users undertaking trips with given

characteristics to the average generalized trip cost and other explanatory variables. This cost is defined, in analogy with Chapter 2, as a (linear) combination of the resources spent by the user on a trip (time, money, stress, etc.) with weights reflecting user's travel behavior. The cost parameters (weights) may vary according to trip purpose and socio-economic category, i.e. user class, and are calibrated together with the demand model, for example in the context of path and mode choice models.

In the following the generalized cost of a trip undertaken between o and d with mode m and following path k by the users of class i will be indicated by $g_{odmk}(i)$. This is equivalent to the cost g_k on path k in mode m network; for uniformity of notation, the zone pair connected by the path and the mode (or mode combination) have been kept explicit.

A simplified specification of the generalized cost analogous to that implicit in the expression (10.4.2) is:

$$g_{odmk}(i) = \beta_1(i)t^i_{odmk} + \beta_2(i)mc^i_{odmk} \qquad (10.4.9)$$

where t and mc are respectively the travel time and the monetary cost. The coefficients here have been explicitly denoted by users class i. Also in this case the generalized cost can be expressed in monetary units by dividing it by the cost coefficient $\beta_2(i)$.

To introduce the calculation method of perceived surplus variations for descriptive demand models, the elementary system consisting of a single O-D pair connected by a single mode and a single path as shown in Fig. 10.4.4 is considered first. Furthermore, it is assumed that all the users belong to one class, i.e. that they have the same behavioral parameters.

In this case the demand model can be formally written as $d_{od} = d_{od}(g_{od})$ which gives the average number of users undertaking a trip for each value of the generalized average cost in the reference period.

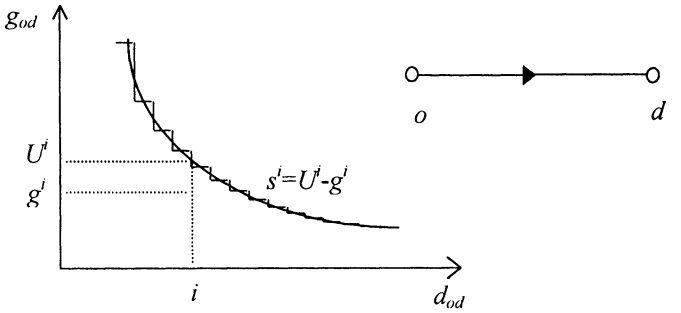


Fig. 10.4.4 – Demand curve of a single O/D pair, mode and path system.

The relationship $d_{od}(g_{od})$ can be represented in the plane and usually has a diagram similar to that described in Fig. 10.4.4. The demand curve, in its traditional

neo-classic interpretation, represents the ordering of individual journeys (or system users) on the basis of the generalized cost they are willing to pay to undertake the journey. In other words, the marginal journey (or user) corresponding to each abscissa has a total trip utility (or willingness to pay) equal to the value of the generalized cost (on the vertical axis). An increase in the cost would discourage this marginal user from making the trip and therefore reduce the value of the demand d_{od}.

Let g_{od}^{NP} be the generalized cost and $d_{od}\left(g_{od}^{NP}\right)$ the number of users traveling in the non-project state. For all journeys undertaken, except the marginal one, there is a net utility, or surplus, given by the difference between the amount they would be willing to pay and the cost that is actually paid (see Fig. 10.4.4). If as a result of project P the generalized cost is reduced to g_{od}^{P}, the number of users traveling increases to $d_{od}\left(g_{od}^{P}\right)$, as described in Fig. 10.4.5.

To calculate the total surplus variation resulting from project P, a distinction should be made between journeys undertaken in the state NP and those undertaken only as a consequence of cost reduction (journey generated by the project)[8]. For the generic journey/user i of the first group, the variation in surplus will be given by:

$$Ds = (U^{i} - g_{od}^{P}) - (U^{i} - g_{od}^{NP}) = g_{od}^{NP} - g_{od}^{P} \qquad (10.4.10)$$

i.e. by the difference in the generalized cost in the states NP and P. The total surplus variation DS'_{p} for all the journeys/users of this group is therefore:

$$DS'_{P} = d_{od}(g_{od}^{NP}) \cdot (g_{od}^{NP} - g_{od}^{P}) \qquad (10.4.11)$$

and is represented by the area A in Fig. 10.4.5.

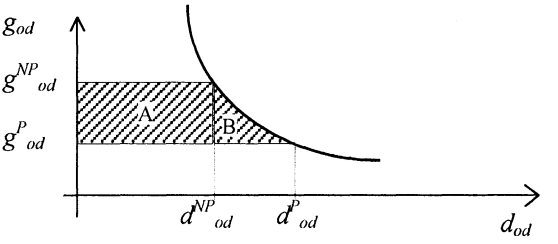


Fig. 10.4.5 – Surplus variations between project (P) and non-project (NP) states: case of cost reduction.

The generic journey i generated by the cost reduction in project P will have a surplus $U^{i} - g^{P}_{od}$, as opposed to a null surplus in the state NP. The total surplus variation for the journeys generated by the project, $d^{*}_{od} = d_{od}(g^{P}_{od}) - d_{od}(g^{NP}_{od})$, will therefore be given by the area B in Fig. 10.4.5. Typically it is assumed that all

generated journeys d^*_{od} have the same utility given by the average value of the interval $[g^{NP}_{od}, g^{P}_{od}]$, i.e. $U^i = (g^{NP}_{od} + g^{P}_{od})/2$, and therefore the total surplus for the generated demand can be calculated as:

$$DS^*_P = d^*_{od}\left[\frac{g^{NP}_{od} + g^{P}_{od}}{2} - g^{P}_{od}\right] = \frac{1}{2} d^*_{od}(g^{NP}_{od} - g^{P}_{od}) \qquad (10.4.12)$$

The total surplus variation will be given by the sum of the terms (10.4.11) and (10.4.12):

$$DS_P = DS'_P + DS^*_P = d_{od}(g^{NP}_{od})(g^{NP}_{od} - g^{P}_{od}) + \frac{1}{2}[d_{od}(g^{P}_{od}) - d_{od}(g^{NP}_{od})](g^{NP}_{od} - g^{P}_{od}) =$$

$$= \frac{1}{2}[d_{od}(g^{P}_{od}) + d_{od}g^{NP}_{od})] \cdot [g^{NP}_{od} - g^{P}_{od}] \qquad (10.4.13)$$

Expression (10.4.13) can be interpreted as the product of the "average" demand between the states *P* and *NP* for the variation in the corresponding generalized cost.

The exact expression of the surplus variation can be obtained by calculating the hatched area in Fig. 10.4.5 as the integral of the demand function $d(g)$:

$$DS_p = -\int_{g^{NP}}^{g^{P}} d(g)dg \qquad (10.4.14)$$

The results described still hold if the project increases the generalized cost ($g^{P}_{od} > g^{NP}_{od}$), as described in Fig. 10.4.6. In this case, clearly there will be a reduction of surplus and a decrease in the number of trips; the surplus variation can also be obtained through the algebraic sum of (10.4.11) and (10.4.12), in this case both negative.

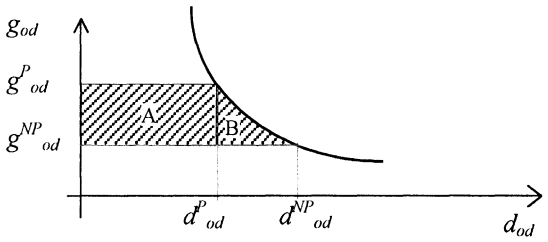


Fig. 10.4.6 – Surplus variations between project (*P*) and non-project (*NP*) states: case of cost increase.

The concept of surplus variation and expressions (10.4.13) and (10.4.14) can be generalized to the case in which there are multiple cost "dimensions" (e.g. multiple

destinations and/or modes and/or paths). However this generalization is neither straightforward not universal. Consider, in fact, a slightly more complex case with two possible alternatives, for example two paths with costs g_1 and g_2, (see Fig. 10.4.7); the two demand curves can be defined as $d_1(g_1\ g_2)$ and $d_2(g_1\ g_2)$. The demand, i.e. the number of trips, on each path depends on the cost of both paths with a diagram similar to that described in Fig. 10.4.7. The demand function can be obtained, for example, combining emission and path choice models. In this case the integral (10.4.14) can be substituted with:

$$DS = - \int_{(g_1^{NP}, g_2^{NP})}^{(g_1^{P}, g_2^{P})} \sum_{i=1,2} d_i(g_1, g_2) dg_1 dg_2 \qquad (10.4.15)$$

whose calculation usually depends on the integration path followed.[9]

For the calculation of the surplus variation, two heuristic approaches can be followed, corresponding to two approximate methods for the calculation of integral (10.4.15).

The first approach, which can be defined as *average demand*, calculates the surplus variation as:

$$DS_P = \frac{1}{2} \sum_{i=1,2} (d_i^{NP} + d_i^{P})(g_i^{NP} - g_i^{P}) \qquad (10.4.16)$$

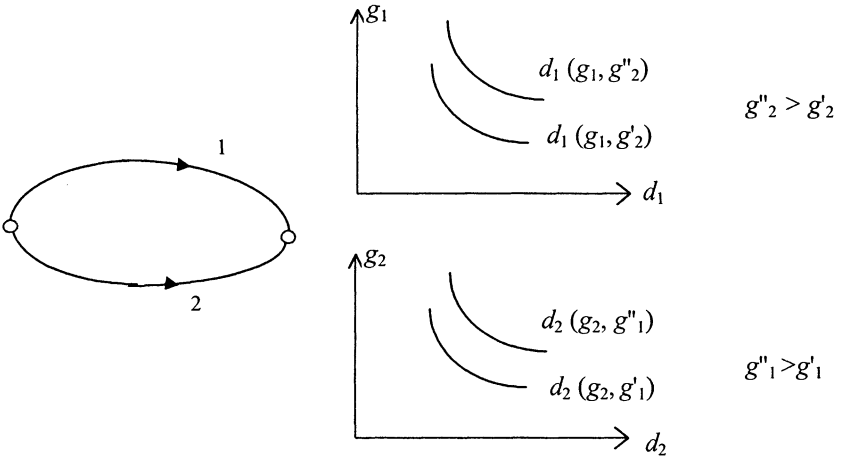


Fig. 10.4.7 – Demand curves for a system with two paths.

where d_i^{P} and d_i^{NP} are respectively equal to $d_i(g_1^{P}, g_2^{P})$ and $d_i(g_1^{NP}, g_2^{NP})$. The expression (10.4.16) can be interpreted as the summation extended to all the

dimensions taken into consideration (in this case, the two paths) of the product of the average demand between the states P and NP, and the cost variation relative to that dimension. Expression (10.4.16) corresponds to the sum of the two hatched areas in Fig. 10.4.8.

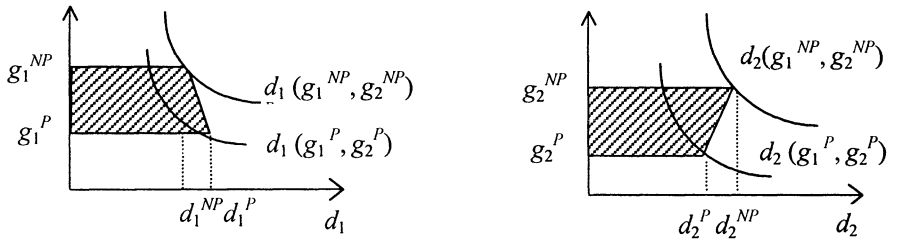


Fig. 10.4.8 – Calculation of the surplus variation with the average demand method.

The alternative approach, which can be defined as *average cost*, reduces the problem to the case of a single choice dimension considering an average trip cost $\bar{g}$ given by the weighted average of the costs relative to each dimension:

$$\bar{g}^{P} = p_1(g_1^{P}, g_2^{P})g_1^{P} + p_2(g_1^{P}, g_2^{P})g_2^{P} \qquad (10.4.17)$$

$$\bar{g}^{NP} = p_1(g_1^{NP}, g_2^{NP})g_1^{NP} + p_2(g_1^{NP}, g_2^{NP})g_2^{NP}$$

where p_1 and p_2 are the demand shares of each dimension, $p_i=d_i/(d_1+d_2)$. In this approach the demand curve represents the diagram of the total demand $d^T = d_1 + d_2$ as the average cost $\bar{g}$ varies (see Fig. 10.4.9). The surplus variation can therefore be calculated by using the expression (10.4.13):

$$DS_P = \frac{1}{2}\left[d^T\left(\boldsymbol{g}^{P}\right) + d^T\left(\boldsymbol{g}^{NP}\right)\right]\left(\bar{g}^{NP} - \bar{g}^{P}\right) \qquad (10.4.18)$$

The surplus variation expressed by (10.4.18) can be interpreted intuitively as the product of the average of the total demand between the states P and NP and the variation of average cost between the two states. Comparing expressions (10.4.16) and (10.4.18), it can be deduced immediately that the two approaches give different results as can be verified from Fig. 10.4.10.

In the general case, the partial share demand model can be expressed conveniently as the product of the demand level and the fraction of trips with given characteristics:

$$d^i_{odmk} = d^i_{o.}\left(\boldsymbol{SE}^i\boldsymbol{g}^i\right)p^i_{dmk\,/\,o}\left(\boldsymbol{SE}^i\boldsymbol{g}^i\right) \qquad (10.4.19)$$

where $d^i_{o.}$ is the number of trips from the zone o undertaken by users of the segment i and $p^i_{dmk/o}$ is the fraction of these trips with the characteristics (dmk).

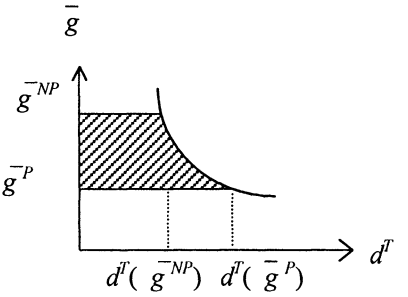


Fig. 10.4.9 – Total demand curve as a function of the average trip cost.

As stated in Chapters 4 and 5, both d^i_o and $p^i_{dmk/o}$ depend on a vector of socio-economic and activity system attributes, ***SE***, as well as on a vector of level-of-service attributes, expressed by the perceived generalized costs for all destinations, with all modes and all paths $\boldsymbol{g}^i$. For simplicity of notation, the dependence on the variables ***SE*** will be implied. The surplus variation corresponding to the passage from state NP with costs $\boldsymbol{g}^{NPi}$ to state P with costs $\boldsymbol{g}^{Pi}$ for users' class i, can be calculated by extending the two previous approximate expressions to the general case. The average demand method, expressed by (10.4.16), therefore yields:

$$DS_p(o,i) = \frac{1}{2}\Sigma_{dmk}\left[d^i_{odmk}(\boldsymbol{g}^{NPi}) + d^i_{odmk}(\boldsymbol{g}^{Pi})\right]\cdot\left(g^{NPi}_{odmk} - g^{Pi}_{odmk}\right) \qquad (10.4.20)$$

On the other hand, the average cost method, expressed by (10.4.18) yields:

$$DS_p(o,i) = \frac{1}{2}\left[d^i_{o.}(\boldsymbol{g}^{Pi}) + d^i_{o.}(\boldsymbol{g}^{NPi})\right]\cdot\left(\bar{g}^{NPi} - \bar{g}^{Pi}\right) \qquad (10.4.21)$$

with

$$\bar{g}^{Pi} = \Sigma_{dmk}\, p^i_{dmk/o}(\boldsymbol{g}^{Pi})\, g^{Pi}_{odmk}$$

and

$$\bar{g}^{NPi} = \Sigma_{dmk}\, p^i_{dmk/o}(\boldsymbol{g}^{NPi})\, g^{NPi}_{odmk}$$

Expressions (10.4.20) and (10.4.21) are the equivalent of expression (10.4.6) in the case of descriptive demand models. The calculation of the surplus variations for all system users can be carried out by adding expressions (10.4.20) or (10.4.21) for

all classes, all zones, all trips purposes and all user classes. An example of the calculation of DS_p for the users of a single market segment with two alternative destinations and two modes/paths for each destination is described in Fig. 10.4.10. However, since surplus variations brought about by project P may be positive for some classes of users, zones or phases of the project and negative for others, it should be recommended to differentiate these values as for behavioral demand models.

c) Comparison between calculation methods

Variation of perceived net utility (surplus) for system users can be calculated either following the behavioral interpretation of random utility models or treating the model as a descriptive demand function. In the second case, the exact calculation poses some definition problems and two different simplified approaches have been proposed. The behavioral approach is certainly more consistent and elegant since it is based on an explicit theory of behavior. It also has two further application advantages[10].

The first advantage stems from the possibility of taking into account surplus variations from variations of attributes, traditionally not considered as components of the generalized cost. In this way surplus variations from increases in the availability of transportation services (e.g. new connections), in travel comfort, or from the provision of information to users, can be evaluated. This obviously requires that these variables are included as explicit or implicit attributes (e.g. alternative specific constants) in systematic utility functions.

On the other hand these effects could not be assessed with the descriptive method since they do not correspond to a reduction in the generalized cost (usually made up of "negative" attributes such as time, monetary cost, etc.). Some paradoxical results could be obtained when increased demand for a mode of superior "quality", but of greater generalized cost, yields a negative surplus variation, i.e. a disbenefit for the users.

The other advantage arises from the possibility of computing surplus variations corresponding to the introduction of alternatives not available in the non-project state, avoiding obvious paradoxes. This point can be clarified with the example in Fig. 10.4.11. The system in the *NP* state offers a single alternative (e.g. a single path) for the single pair (o,d). In the P state a second path with an higher "generalized cost" is added; the total demand is assumed to be constant and the distribution between the two paths is obtained with the Binomial Logit model described in the figure. The surplus variation with the behavioral method can be calculated substituting the logsum variable (10.4.4) into total average surplus (10.4.6). Since the Expected Maximum Perceived Utility function is monotone increasing with respect to the number of available alternatives (see section 3.5), the surplus of users in the state P increases with respect to the state *NP*. Vice-versa, calculation of surplus variation with descriptive methods poses some problems. First, the average demand method corresponding to the expression (10.4.20) cannot be used since it is not possible to define a cost g_2^{NP} for the new path. The average

cost method corresponding to the expression (10.4.21) can be used since it requires only the total demand d^{NP} and d^{P} and the weighted average of path costs for the states P and NP can be computed. However, because of the increase in the average cost, the method gives a negative surplus variation, i.e. a reduction in the net utility for the system's users. This outcome is clearly paradoxical since an increase in supply should correspond to an increase in users' surplus if some users are using the new path. The explanation is to be found in the difference between the assumptions underlying the demand model and the calculation of the surplus.

The Logit model, beyond its behavioral interpretation, assigns a positive probability to alternatives with greater generalized cost, implying that the cost perceived by the users is different from the average. Vice-versa the average cost method associates to the users the "objective" average cost of the alternative chosen.

The two methods would give the same outcome only in the case of deterministic utility choice model. In this case, in fact, the whole demand would choose path 1 also in the state P, the average cost would be equal to g_1 and the surplus variation would be equal to zero.

This result can be generalized since, as discussed in section 3.5, the EMPU variable for deterministic choice coincides with the maximum utility (minimum cost) value. However, if the demand model were a deterministic utility one, it would be a behavioral model and the previous behavioral case would apply.

The surplus variation can also be calculated by using a mixed approach in which the "average cost" descriptive method (10.4.21) is applied by substituting the average costs $\bar{g}^{P}$ and $\bar{g}^{NP}$ with the corresponding EMPU values s^{P} and s^{NP} calculated on the choice dimensions for which a behavioral model is used.

For example in the case of Multinomial Logit model on three dimensions $d\ m\ k$ and parameter $\theta = 1$ it would result:

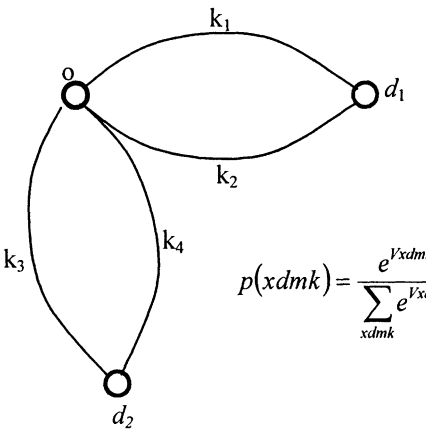
$$s_P(o) = ln\ \Sigma_{d'm'k'}\ exp\ [V_{d'm'k'}] = Y_{o.}$$

which is the *accessibility* from zone o to all destinations with all the available modes and paths.

In this case, the curve expressing the demand level $d_o(s)$ as a function of the EMPU variable on the choice dimensions $d\ m\ k$ can be interpreted as the ordering of the trips (journeys) with respect to the corresponding *average perceived net utility*. The number of users undertaking a trip increases with s. The diagram of the demand function with respect to the inclusive utility and the area corresponding to the surplus variation for an increase in the EMPU s is shown in Fig. 10.4.12[11]. Also in this case a linear approximation can be used for the calculation of surplus variation:

$$DS = 1/2\ (d^{P} + d^{NP})(s^{P} - s^{NP})$$

bearing in mind that EMPU and cost have opposite signs.



$$p(xdmk) = \frac{e^{Vxdmk}}{\sum_{xdmk} e^{Vxdmk}}$$

Non project (NP)	$g^{NP}_{od_1k_1} = 1{,}2\ d_{od_1k_1}$	$(\boldsymbol{g}^{NP}) = 100 \cdot p^{NP}_{od_1k_1} = 36$	
	$g^{NP}_{od_1k_2} = 1{,}4\ d_{od_1k_2}$	$(\boldsymbol{g}^{NP}) = 100 \cdot p^{NP}_{od_1k_2} = 30$	$d_o\ (\boldsymbol{g}^{NP}) = 100$
	$g^{NP}_{od_2k_3} = 2{,}0\ d_{od_2k_3}$	$(\boldsymbol{g}^{NP}) = 100 \cdot p^{NP}_{od_2k_3} = 17$	
	$g^{NP}_{od_2k_4} = 2{,}0\ d_{od_2k_4}$	$(\boldsymbol{g}^{NP}) = 100 \cdot p^{NP}_{od_2k_4} = 17$	
Project (P)	$g^{P}_{od_1k_1} = 1{,}0\ d_{od_1k_1}$	$(\boldsymbol{g}^{P}) = 100 \cdot p^{P}_{od_1k_1} = 29$	
	$g^{P}_{od_1k_2} = 1{,}0\ d_{od_1k_2}$	$(\boldsymbol{g}^{P}) = 100 \cdot p^{P}_{od_1k_2} = 29$	$d_o\ (\boldsymbol{g}^{P}) = 100$
	$g^{P}_{od_2k_3} = 1{,}2\ d_{od_2k_3}$	$(\boldsymbol{g}^{P}) = 100 \cdot p^{P}_{od_2k_3} = 23$	
	$g^{P}_{od_2k_4} = 1{,}4\ d_{od_2k_4}$	$(\boldsymbol{g}^{P}) = 100 \cdot p^{P}_{od_2k_4} = 19$	

Average demand method

$$DS_P(o) = \frac{1}{2} \sum_{dmk} \left[d_{odmk}(\boldsymbol{g}^{NP}) + d_{odmk}(\boldsymbol{g}^{P}) \right] \cdot \left(g^{NP}_{odmk} - g^{P}_{odmk} \right) =$$

$$= 0.5 \cdot \left[(36+29) \cdot 0.20 + (30+29) \cdot 0.4 + (17+23) \cdot 0.8 + (17+19) \cdot 0.6 \right] = 45.1$$

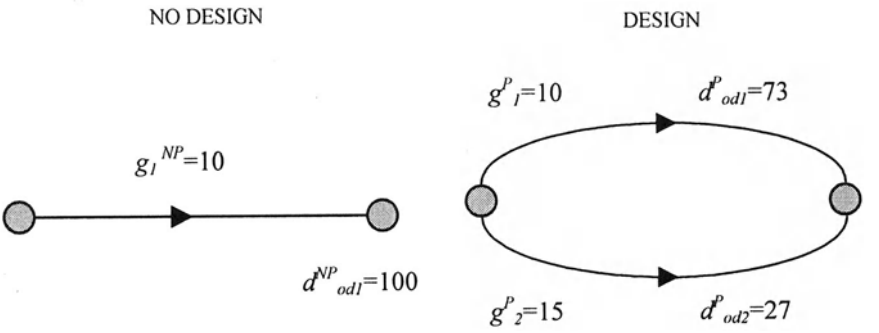
Average cost method

$$DS_P(o) = \frac{1}{2} \left[d_{o.}(\boldsymbol{g}^{P}) + d_{o.}(\boldsymbol{g}^{NP}) \right] \cdot \left(\bar{g}^{NP} - \bar{g}^{P} \right) =$$

$$= 0.5 \cdot (100+100) \cdot \left[(1.2 \cdot 0.36 + 1.4 \cdot 0.30 + 2.0 \cdot 0.17 + 2.0 \cdot 0.17) + \right.$$

$$\left. - (1.0 \cdot 0.29 + 1.0 \cdot 0.29 + 1.2 \cdot 0.23 + 1.4 \cdot 0.19) \right] = 41.0$$

Fig. 10.4.10 –Calculation of surplus variation with the average demand and average cost methods for the system of Fig. 10.4.3.



DESCRIPTIVE APPROACH

$$DS = 100 \cdot [10 - (0.73 \cdot 10 + 0.27 \cdot 15)] = -135$$

BEHAVIORAL APPROACH

$$V_k = -0.2 \cdot g_k$$

$$S^{NP} = 100\, ln[exp(-0.2 \cdot 10)] = -200$$

$$S^{P} = 100\, ln[exp(-0.2 \cdot 10) + exp(-0.2 \cdot 15)] = -169$$

$$DS = S^P - S^{NP} = 31$$

Fig. 10.4.11 – Calculation of the surplus variation following descriptive and behavioral approaches.

From the previous discussion it follows that surplus variation, as far as is possible, should be computed on the basis of Expected Maximum Perceived Utility variables, especially when the project increases the number of available alternatives[12] or at least with a mixed approach using EMPU variables on the choice dimensions more closely connected to the changes on the transportation system, such as mode and path choice.

10.5. Methods for the comparison of alternative projects

There are several methods for comparing alternative projects of transportation systems. This section will shortly present the quantitative methods which are most used in applications for economic evaluation, namely the traditional Benefit-Cost analysis and some Multi-Criteria analysis methods. The reader is referred to the vast literature on the subject for further information.

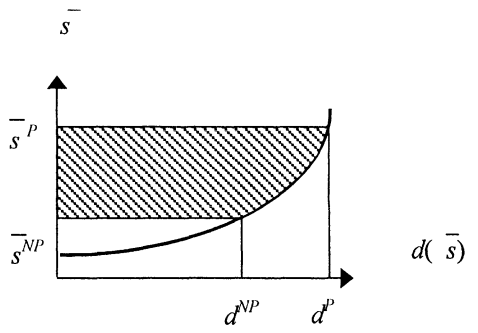


Fig. 10.4.12 – Demand function with respect to EMPU values.

10.5.1. Benefit-Cost analysis

Benefit-Cost (*B/C*) analysis compares alternative projects considering their effects expressed in monetary units. A single economic aggregate is formed in which different impacts are algebraically summed, considering with positive sign (Benefits) "income" items and with negative sign (Costs) "disbenefit" items. Benefits and costs are obviously related to the subject for whom the analysis is performed.

Applications of the *B/C* method from the viewpoint of a single public decision-maker (typically a governmental agency) can consider for each year t of the economic life of project "P_i" all or some of the following effects:

CC difference between the construction cost of the project and the construction and extra-ordinary maintenance costs of non-project, if any. It should be remembered that investment already decided could be included in the *NP* state. In some applications, a negative construction cost *CC* (i.e. a benefit) is assumed for the final year; this corresponds to the residual value of the project at the end of the analysis period. In this way it is possible to reduce the unavoidable arbitrariness in the definition of the technical life of the project.

CVT difference between investment costs in vehicles and technologies for the project and non-project states. Also in this case the *NP* state might require investment in means of production.

CMO difference between maintenance and operation cost for project and non-project states.

REV difference between direct (sale of transportation services) and indirect

(commercial activities) revenues in the project and non-project states.

TR difference between taxes and duties revenues in the project and non-project states.

DS variation of surplus perceived by the users of the transportation system in the project and non-project states expressed in monetary units. This is obtained by adding up the variations of perceived surplus, for different user classes.

UNPB variations in benefits not perceived by the users between the project and non-project states. These benefits might include variations of costs due to accidents, vehicle consumption (lubricants, tires, etc.) and other non out-of-pocket costs not perceived by the users in their travel-related choices. All these benefits are expressed in monetary units; the variable has a positive sign if there is a reduction in these costs.

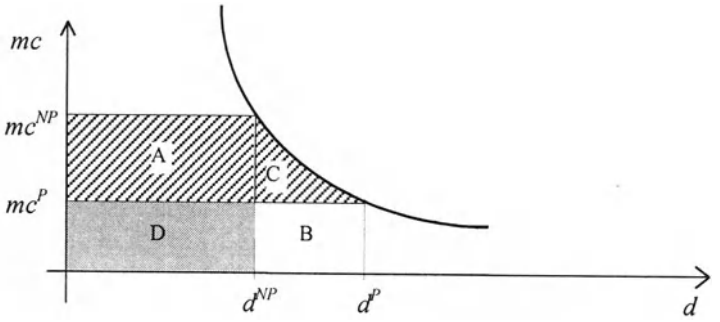
NUI variation of the impacts for non-users between the project and non-project states. Impacts on the environment (e.g. reduction of pollutants emission appropriately expressed in monetary terms) and on the economical and territorial system, as described previously, can be included in this variable. The variable is sometimes indicated as indirect benefits and is positive if these benefits increase.

The above variables are usually calculated with market prices, when available, possibly reduced by the transfers internal to the Public Administration (VAT, income and fuel taxes, etc,). For example, construction, maintenance, and operation costs can be computed by evaluating the resources employed at market prices minus VAT and other taxes. In some applications, market prices are replaced by *shadow prices*, or *opportunity costs*, which reflect the value of a particular resource to the community. Shadow prices can be assigned when there is no market price or when this is modified to take into account objectives or constraints of social interest not reflected by market mechanisms. The opportunity cost of labor, for example, might be lower than the market price of manpower when there is a high level of unemployment, and its reduction is one of the objectives of the project. In this case, the opportunity cost could be obtained as the difference between net market price and the unemployment subsidy for each category of workers.

It is important to stress that the variables considered and the way they are computed both depend on the viewpoint from which B/C analysis is performed. Also in the case of a public operator (public sector analysis) there may be different viewpoints. For example construction costs may be reduced by non-reimbursable government grants for Local Administrations.

Whatever the point of view of the evaluation, double counts of the same effect with the same sign in several variables must be avoided. Some effects may be present with different signs in two or more variables; for example the fares paid by the users can be counted in traffic revenues with a positive sign (benefit) and in the perceived surplus variation with a negative sign (cost). The same occurs with

gasoline taxes, and other variables. Effects of this kind could even be excluded from the *B/C* analysis, as proposed by some analysts for traffic revenues. Their exclusion, however, is acceptable only in the special case in which the effects count linearly in all terms. In the previous example, this is the case if the monetary cost appears linearly in generalized user costs and if surplus variation is computed through descriptive methods (10.4.20) and (10.4.21) with rigid demand. However, if this variable appears non-linearly in different terms, for example the monetary cost is used in the EMPU variable (10.4.4) and (10.4.6) for the evaluation of user's surplus variation, and/or there are variations in the level of demand, it must necessarily be accounted for twice. Fig. 10.5.1 shows graphically the difference between surplus and revenues variations in the case of generalized cost coinciding with monetary cost; the two variations would coincide in absolute value only in the special case of rigid demand (areas B and C equal to zero).



$$\Delta_{rev.} = d_P \cdot mc^P - d_{NP} \cdot mc^{NP} = (B+D)-(A+D) = B-A$$

$$\Delta_{surplus} = A+C$$

$$|\Delta_{rev.}| \neq |\Delta_{surplus}|$$

Fig. 10.5.1 – Difference between surplus and revenues variations

An example of potential double counting is given by the case of increased accessibility (reduction of the generalized transportation cost) of one zone compared with others. This effect usually leads to an increase in the real estate values in the zone as a consequence of residents' and/or firms' willingness to pay for the greater accessibility. If the surplus variations for the users and the variations in real estate values were both counted as benefits, the accessibility effect of the project would be accounted for twice and, in this particular case, the overall benefits would be overestimated. In this example the variation of real estate values should not be considered, or it should be accounted for with opposite signs for those who benefit from them (i.e. the landlords) and those who incur penalties (i.e. renters or buyers).

It is clear that these effects will be very different for the different categories and their distribution within the society as a whole cannot be considered irrelevant.

Once the relevant effects have been defined and measured in monetary units, different alternative projects are compared using synthetic indicators or aggregate variables. Benefits and costs relative to different years are compared by means of the interest or discount rate r. This is defined as the relative increase of the sum M after one year:

$$r = \frac{M^{t+1} - M^t}{M^t}$$

$$M^{t+1} = (1 + r)\, M^t$$

The value M^t, of a sum M_o available today after t years can therefore be calculated as:

$$M^t = M_o\,(1 + r)^t$$

from which it follows that the present value M_o of a sum M^t spent or gained after t years is:

$$M_o = \frac{M^t}{(1+r)^t} \qquad (10.5.1)$$

Several synthetic indicators have been proposed for the comparison between benefit and cost flows for the different projects P_i. The *Net Present Value* (*NPV*) brings to the present the effects calculated for a period of T years assuming a constant discount rate r:

$$NPV_i(r) = \sum_{t=1}^{T} \frac{(DS_i^t + UNPB_i^t + NUI_i^t + TR_i^t + REV_i^t - CC_i^t - CVT_i^t - CMO_i^t)}{(1+r)^t} \qquad (10.5.2)$$

The *Internal Return Rate* (*IRR*) is defined as the value of the discount rate r_o such that the *NPV* calculated over a period of T years is equal to zero:

$$IRR_i = r_o : NPV_i(r_o) = 0 \qquad (10.5.3)$$

Using the first indicator, the generic project P_i is preferable to non-project *NP* if its *NPV* is positive; the project P_i is preferable to the project P_j if $NPV_i > NPV_j$. The superiority of a project P_i over P_j may depend significantly on the discount rate r used for the calculation of *NPV*, as shown in Fig. 10.5.2. Projects with lower investment costs and fewer benefits usually are positively affected by higher values of r (project P_B in fig. 10.5.2), while low discount rates "favor" more costly projects

with greater future benefits (project P_A in fig. 10.5.2). Higher discount rates, in fact, reduce the present value of project benefits, usually obtained some years after the investment is made; on the contrary, project investment costs, borne in the early years, are less sensitive to discount rates.

A project P_i is preferable to the non-project NP in terms of Internal Return Rate if its IRR is above the "social" discount rate and is preferable to the project P_j if $IRR_i > IRR_j$. The discount rate used to compute NPV or to compare the IRR can be selected in several different ways. One possibility is the interest rate prevailing in the economic system of analysis. Other more complex methods adopt the social opportunity cost of the capital based on the returns potentially achieved with alternative uses, the social marginal utility of consumption or measure of risk connected to the project. This subject has been discussed at length in the economic literature to which the interested reader is referred. Here it is only worth mentioning that the discount rate has important implications of value for present consumption compared with future effects; these should be explicitly stated and may depend on the point of view of the analysis.

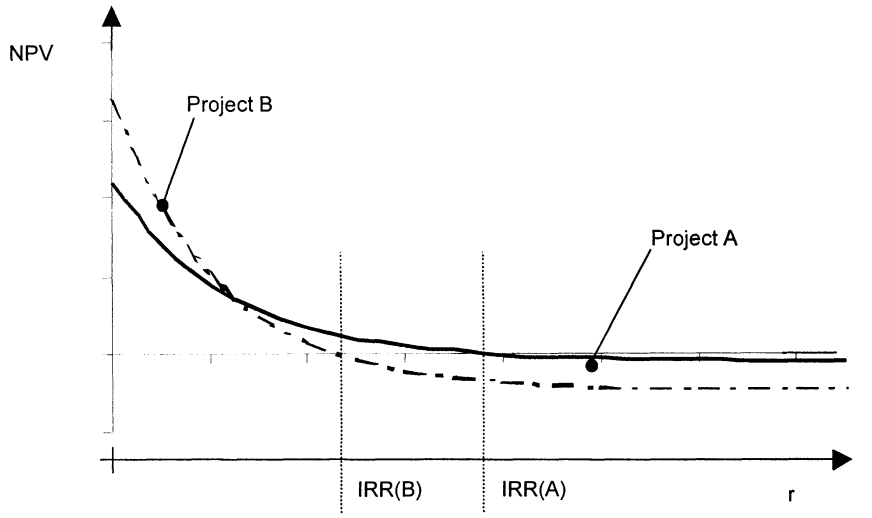


Fig. 10.5.2 – Net Present Value as a function of the discount rate.

Benefit-Cost analysis has undergone some criticism both in terms of the ways it is used and its theoretical foundations.

Some objections of the first type can be summarized as follows.

i) The use of market prices as indicators of the "social" value of resources is theoretically correct only under the critical assumptions of socially optimal income distribution and of perfectly competitive markets. In reality, both assumptions are almost always far from the truth and the use of market prices implies judgements on

income distribution and is inconsistent with the results of the welfare economy inspiring it. In alternative, as indicated, shadow prices can be used, if "social" objectives or constraints are pursued through the project. For example, objectives such as reduction of unemployment, reduction of air pollution or energy consumption can be reflected in shadow prices of labor and fuel. The rigorous calculation of shadow prices is extremely complex, and in practice rough estimates are often used.

ii) The evaluation of the effects for users and non-users of the system may be incomplete or inexact. In some applications, impacts are computed only for the users of the planned facilities, ignoring the effects on the remaining part of the transportation system. While this approximation may be acceptable in some cases, in others it may significantly distort the results of the analysis. It is quite common, in fact, that because of the interdependencies in a transportation system, the effects on the travelers not using the new facilities or services directly are comparable to those on direct users. This and other similar criticisms can be overcome by analyzing the transportation system as described in previous sections.

The "structural" criticisms of Benefit-Cost analysis relate to aspects that cannot be eliminated by a proper application of the procedure.

i) The aggregation of the effects on different groups, once again implies value judgements on the "optimality" of the present income distribution and on the indifference with respect to the income redistribution that may be caused by the project. For example, it is assumed that an increase in generalized cost for some users is compensated by a reduction of the same magnitude of the cost to other users. Furthermore, the perceived cost depends on the user's income, thus variations in travel time of the same amount produce larger perceived surplus variations for higher-income groups and so on.

ii) If the individuals receiving benefits from the project would pay for such benefits in monetary terms and this sum exceeded that necessary to compensate those receiving negative effects, the project would result in a net increase in public "welfare". In reality, compensations are often only hypothetical. On the other hand, in the *B/C* analysis a project is considered socially preferable to another if the potential willingness to pay of those who benefit is superior to the amount needed to compensate those receiving a damage, regardless of the actual occurrence of these transactions.

iii) The Benefit/Cost analysis is limited to effects that are, or can be, expressed in monetary units, ignoring a number of effects which cannot be significantly measured in monetary units as discussed in section 10.2. This implicitly privileges the objective of economic efficiency over other social and environmental objectives.

On the basis of these considerations, many economists agree on assigning to the *B/C* analysis, from the point of view of the public decision-maker, a role that is essentially normative and/or conventional. In the former case the main elements of the analysis, i.e. effects, prices, discount rates, etc. are fixed by the agencies funding public projects in order to receive comparable and homogenous proposals to be compared. In the latter, the parameters of *B/C* analysis are consolidated from

practical applications in specific sectors. Alternatively, *B/C* analysis should not be considered as a comprehensive evaluation method, but rather as an evaluation of economic impacts for some actors of the decisional process, taking into account only monetary or monetarily quantifiable "costs" and "benefits". In this interpretation *B/C* analysis can be considered as a synthetic way for evaluating the impact on economic efficiency for the several possible actors interested. Thus several *B/C* indicators could be computed representing different actors such as users, service operators, public agencies; these indicators can be used together with others in the context of a wider Multi-Criteria analysis discussed below.

10.5.2. Multi-Criteria analysis

As stated in the previous sections, transportation system projects may induce effects of different types and decision-makers generally have multiple goals, which may conflict with each other. Each effect described in section 10.2 corresponds to an impact on one or several actors and, as such, can be transformed into an objective. Thus, increasing user surplus, reducing expenditure, increasing revenues, increasing social equity and accessiblity, increasing the efficiency of the transportation system, reducing environmental impacts, and increasing the economic efficiency of the system are all possible objectives for the same decision-maker and/or for the actors involved in the decision process. As stated these objectives often conflict with each other; the maximization of user surplus might, for example, conflict with the reduction of noise and air pollution and with the minimization of capital investments.

Multi-Criteria (*MC*) or Multi-Objective (*MO*) analysis aims at supporting the decision-makers to reach a feasible compromise between the different objectives. Applications of Multi-Criteria analysis to public choices, and therefore to the evaluation of transportation projects, has increased significantly in recent years, also in connection with the increase awareness of the limitations of Benefit-Cost analysis. Multi-Criteria analysis is a general term including several techniques differing in theoretical basis, calculation methodologies and fields of application. Only some of these techniques will be outlined in the following to give the reader an idea of this approach.

The different objectives of the decision-makers are first transformed into *evaluation criteria* or *performance indicators*, i.e. quantitative and qualitative variables measuring the level of achievement of the generic objective. For example, the performance indicator corresponding to the objective of increasing users' utility might be the difference between the total surplus of the users. Values of *NPV* and *IRR* may correspond to the objective of increasing the economic efficiency, the indicator corresponding to the objective of reducing air pollution might be the variation of total pollutant emissions and so on. Criteria expressed qualitatively (e.g. with adverbs such as little, much, etc.) can be transformed into quantitative variables by indirect quantitative determination techniques. The following, therefore, refers exclusively to quantitative criteria. The identification of the objectives and their relative evaluation criteria is a crucial phase of any Multi-Criteria evaluation. As a

matter of fact objectives and criteria should be specified with similar levels of detail to avoid distortions in the results of the analysis. The practical rule is to use a balanced number of criteria for the different macro-objectives of the project under study.

A weight $w_m \geq 0$ can be attributed to each criterion m, measuring the importance of the objective corresponding to the criterion m compared with the other objectives for the decision-maker. Obviously, in the definition of weights decision-makers are asked to express value judgements. In principle different sets of weights can be associated to the same set of objectives and their indicators expressing the point of view of different actors in the decision process.

Many methods have been proposed to estimate the unknown weights for each decision-maker as well as to reach a compromise among several decision-makers. The most direct approach, known as the DELFI method, consists of having each decision-maker express separately and explicitly the weight relative to each macro-objective or criterion. Subsequently the interviews are repeated, telling each interviewee the weights stated by the other decision-makers until a compromise is reached. When the weights cannot be obtained directly from decision-makers, other procedures can be used. For example, it is possible to estimate the weights, which would justify the choices made in similar contexts for projects of the same type and size. Following a different approach the decision-maker is asked to express preferences between pairs of alternative hypothetical projects with trade-offs between the different objectives; the implicit set of weights can be estimated as to reproduce as closely as possible the stated choices.[13]

In many *Multi-Criteria* techniques, performance indicators are first processed to allow their comparison. Suppose that M evaluation criteria corresponding to the objectives of the project have been identified and that the value of the m-th performance indicator for the j-th project is expressed through the variable x_{mj}. Variables x_{mj} are usually expressed on a scale increasing with the level of satisfaction. When x_{mj} measures a "negative" effect, e.g. the quantity of emitted pollutants, it can be substituted with the reduction with respect to the maximum value assumed by the indicator:

$$x'_{mj} = (\max_k x_{mk}) - x_{mj}$$

In order to avoid distortions deriving from the use of different scale factors for different indicators the values of the indicators are sometimes substituted by normalized values l_{mj}, included in the interval [0,1]. Several forms of normalization are possible. Some normalization equations are linear in the indicator x_{mj}:

$$l_{mj} = \frac{x_{mj} - \min_k x_{mk}}{\max_k x_{mk} - \min_k x_{mk}}$$

$$l_{mj} = \frac{x_{mj}}{\max_k x_{mk}} \qquad (10.5.4)$$

$$l_{mj} = \frac{x_{mj}}{\sum_k x_{mk}}$$

In some applications indicators are normalized through non-linear, monotone increasing functions, usually denominated utility functions. With utility functions, one can account for the decreasing marginal utility of increasing levels of a given criterion (see Fig. 10.5.3); in this case performance indicators are known as utility indicators and denoted by u_{mj}.

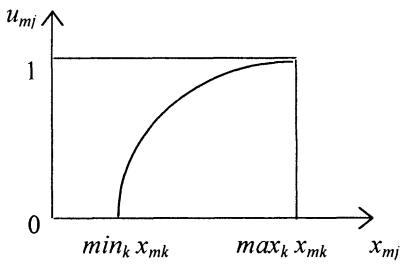


Fig. 10.5..3 – Utility function for a generic performance indicator.

The *evaluation matrix* or *impact tableau* consists of the evaluation indicators (x_{mj} , $x'_{mj,}$ l_{mj} or u_{mj}). This matrix has a number of rows equal to the number of evaluation criteria and a number of columns equal to the number of alternative projects. An example of such a matrix is described in Fig. 10.5.4.

The project j is *dominated* if there exists at least one project h satisfying all the objectives, better than, or at least equal to, project j:

$$x_{mj} \leq x_{mh} \; \forall m = 1,...M \qquad (10.5.5)$$

with at least one of the inequalities (10.5.5) holding as a strict inequality[14].

A non-dominated project is also called efficient. The set of non-dominated projects satisfying the constraints (e.g. budget constraints) is called the project efficiency boundary.

Evaluation criteria	Project 1	2	...	j	...	J	Weights
1	x_{11}	x_{12}	....	x_{1j}	....	x_{1J}	w_1
2	x_{21}	x_{22}	....	x_{2j}	....	x_{2J}	w_2
....	....	....	....	....	....	....	....
m	x_{m1}	x_{m2}	....	x_{mj}	....	x_{mJ}	w_m
....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....
M	x_{M1}	x_{M2}	....	x_{Mj}	....	x_{MJ}	w_M

Fig. 10.5.4 – Evaluation matrix of *J* alternative projects with respect to *M* criteria.

It can be shown that all the points of this boundary are potentially optimal solutions to the decision problem for an appropriate set of weights for the different objectives/criteria.

Multi-Criteria analysis techniques proposed in the literature generate a set of non-dominated solutions (projects) and assist the decision-maker in selecting a reasonable compromise between contrasting objectives. Some of these techniques will be described below using the performance indicators x_{mj}, these can be substituted by l_{mj} or u_{mj} .

Some techniques generate a "continuous" set of non-dominated projects defined by continuous decision variables with explicit relationships, preferably linear, between these variables and their effects[(15)]. In the case of transportation system projects, these conditions are rarely met because of the discrete nature of many projects (new infrastructures, for example), the intrinsic non-linearity of the system (cost functions and demand models) and the complexity of the relationships between control variables and effects (e.g. variations of equilibrium flows and costs following a transportation network project). For these reasons, and to simplify the treatment of the subject, in what follows, reference will be made to the case of a discrete set of *J* alternatives (projects). Furthermore, alternative projects are assumed to be non-dominated since dominated ones on the assumption of monotonicity of preferences, could never be optimal choices under any set of weights.

The role of the analyst and the decision-maker varies greatly with the different techniques proposed. According to some authors, the analyst's task should end after presenting the decision-maker with the list of non-dominated projects together with the available information, processed in such a way to facilitate their understanding to non-specialists. Other methods assume a certain amount of data processing and interaction with the decision-maker.

One approach, known as the *distance method*, identifies the best compromise solution as the project *j* minimizing the "distance" from the *ideal solution*. The latter can be defined as the hypothetical project satisfying all the objectives at the maximum level; the ideal solution is not among the available options, which would otherwise all be dominated. Denoting by x^*_m the highest level of the performance indicator for criterion *m*, the vector $\boldsymbol{x}^* = (x^*_1, x^*_2, ..x^*_M)^T$ represents the effects of

the ideal project. The distance is measured as the weighted *p-norm*, L_p, of the difference between the two vectors x_j and x^* :

$$L_p(x_j, x^*) = [\Sigma_m w_m(x^*_m - x_{mj})^p]^{1/p} \tag{10.5.6}$$

The definition of a norm $L_p(x_j, x^*)$ in the space of the performance indicators is equivalent to defining a multi-attribute utility function allowing the substitution of satisfaction levels of the different objectives. As p varies, different measures of distance between the vectors x_j and x^* are obtained, in particular for p equal to twice the weighted distance in the Euclidean space is obtained:

$$L_2(x_j, x^*) = [\Sigma_m w_m(x^*_m - x_{mj})^2]^{1/2} \tag{10.5.7}$$

if the weights were equal for all criteria, (10.5.7) would be reduced to the Euclidean distance between the two vectors x_j and x^*. As p tends to infinity, the L_p norm distance is defined exclusively by the distance from the ideal to the most distant indicator:

$$L_\infty(x_j, x^*) = max_m|w_m(x^*_m - x_{mj})| \tag{10.5.8}$$

Figure 10.5.5 illustrates the distance functions from the ideal of three projects as the parameter p varies. It can be observed that for low values of p the project with the largest performance indicator for the greatest number of criteria is preferable, while as p increases, the project minimizing the maximum deviation from the ideal project becomes preferable.

Another approach is based on the pairwise comparisons of alternative projects. Methods of the *ELECTRE family* are among the most popular examples of this approach. The most recent version, *ELECTRE IV*, defines the *index of concordance* c_{ij} of the project i compared with project j as a standardized measure of prevalence, or preferability, of i compared with j, the index is equal to one if the project i dominates project j:

$$c_{ij} = \frac{\sum_{m \in S_{ij}} w_m}{\sum_{n=1...M} w_n} \tag{10.5.9}$$

where $S_{ij} \equiv \{m : l_{mi} \geq l_{mj}\}$ is the set of criteria for which the project i is superior, or not inferior to j and l_{mj} is the normalized performance indicator defined by (10.5.4). Obviously, the index c_{ij} will be the closest to one, i.e. project i will be more preferable to project j, the greater the weights w_m of the criteria for which i is superior to j.

The *discordance index* d_{ij} of project i compared to project j is a standardized measure of the "inferiority" of i compared to j. It is equal to one if the maximum

weighted deviation in favor of j among all the criteria for which j is superior, coincides with the maximum absolute weighted deviation between i and j for all the criteria. In formal terms, it results:

$$d_{ij} = \frac{\max_{m \in I_{ij}} \left[w_m (l_{mj} - l_{mi}) \right]}{\max_{n} \left[w_n \left| l_{ni} - l_{nj} \right| \right]} \tag{10.5.10}$$

where $I_{ij} \equiv \{m: l_{mi} < l_{mj}\}$ is the set of criteria indices which the project i is inferior to j.

Concordance and discordance indices can be used differently to compare available alternatives. The *mobile threshold method* calculates the concordance and discordance indices for all the ordered pairs of alternative projects. Alternatives can be ordered by fixing two thresholds, $\bar{c}$ and $\bar{d}$, with $\bar{c} \leq \bar{d}$ and rejecting all the project pairs (i,j) such that c_{ij} is less than $\bar{c}$ (i.e. pairs for which i is not significantly superior to j) and/or d_{ij} is greater than $\bar{d}$ (i.e. i is clearly inferior to j). For the pairs of alternative projects satisfying both requirements it results:

$$c_{ij} \geq \bar{c} \text{ e } d_{ij} \leq \bar{d}$$

These pairs are considered to give a significant indication of superiority of alternative i over alternative j. If the residual pairs still do not lead to a unique ordering –(for example i is preferable to j, j is preferable to k but k is preferable to i), the values of the thresholds $\bar{c}$ and $\bar{d}$ are modified by increasing the former and reducing the latter until a set of project pairs expressing a unique preference is obtained (see Fig.10.5.6). A different method is based on the calculation of a synthetic indicator of preference or superiority index s_i of alternative i as follows:

$$s_i = \sum_{j \neq i} z_{ij} - \sum_{j \neq i} z_{ji}$$

$$\begin{aligned} z_{ij} &= 1 \qquad \text{if } c_{ij} \geq \bar{c} \text{ and } d_{ij} \leq \bar{d} \\ z_{ij} &= 0 \qquad \text{otherwise} \end{aligned} \tag{10.5.11}$$

The superiority index is equal to the number of projects significantly inferior to i minus the number of projects significantly superior to i. The ordering of the alternatives can be carried out on the basis of the values of the indicator s_i, projects with higher values will be preferred.

Evaluation criteria	Project			Weights	Ideal project
	1	2	3		
1	8	7	10	1	10
2	4	6	6	1	6
3	4	6	6	1	6
4	5	7	7	1	7
5	6	8	4	1	8
6	5	4	7	1	7

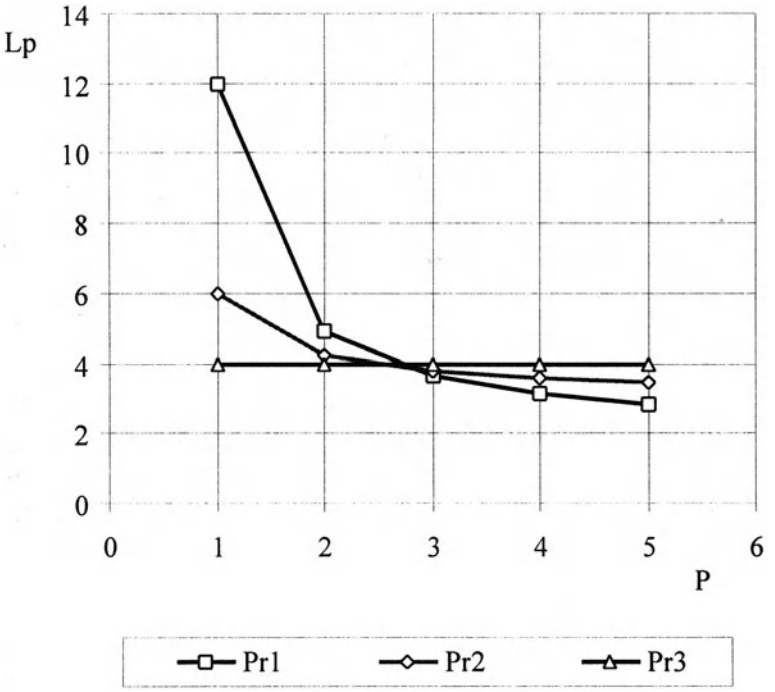


Fig. 10.5.5 – Distance L_P from the ideal project for varying values of the norm *p*.

Alternatives A,B,C,D

$\bar{c}$ =0.30 $\bar{d}$ = 0.70

Pairs of alternatives	Concordance indices	Discordance indices
A-B	0.40	0.20
A-C	0.70	0.50
B-A	0.60	0.50
B-C	0.65	0.30
C-B	0.35	0.60
C-D	0.35	0.40

$\bar{c}$ =0.45 $\bar{d}$ = 0.50

Pairs of alternatives	Concordance indices	Discordance index
A-C	0.70	0.50
B-A	0.60	0.50
B-C	0.65	0.30

Ordering B > A > C

Fig. 10.5.6 – Example of application of the *ELECTRE IV* mobile threshold method.

In Multi-Criteria analysis, whatever the comparison technique, the phase of *sensitivity analysis* is of considerable importance. Sensitivity analysis explores how sensitive the outcome obtained, i.e. the ordering of the alternatives, is to the assumptions on the parameters used. In other words, it attempts to establish whether the solution obtained is stable with respect to variations in the parameters, which, as has been said, are intrinsically arbitrary. As an example, such a technique analyzes the vector of weights w_m in the function (10.5.6) looking for the limits of the values that would not change the ordering of projects. The larger the differences with the adopted weights, the more reliable is the ordering of projects. Sensitivity analyses can be carried out by different methods with different levels of sophistication, the description of these methods, can be found in the specialized literature.

Reference Notes

A clear and concise description of the different approaches to the general problem of planning and public decision making, with special reference to town planning, is given in Alexander (1997) which contains a vast bibliography. Many textbooks deal with the process of transportation planning from different viewpoints. References representing differing, and somehow contrasting positions, are Hutchinson (1974), Manheim (1979) and Meyer and Miller (1984). Wachs (1985), Bianco (1986) and Meyer and Miller (1984) contain a commented bibliography of the theoretical developments of the concept of transportation systems planning. A classification of the different levels of planning is described in Florian, Gaudry and Lardinois (1988).

The economic literature relevant to the welfare theory applied to the analysis of investments is quite substantial and a systematic analysis is well beyond the scope of this book. Among the many texts on the subject reference can be made to the classic book of Mishan (1974).

There is also a vast literature on Multi-Criteria analysis. The fundamentals of these techniques can be found in Chankong and Haimes (1983), Voogd (1983), Haimes and Chankong (1985), Nijkamp et al, (1990). The Electre method is described in Voogd (1983).

Applications relative to the evaluation of transportation system projects are to be found in almost all books on transportation planning. The traditional approach of Benefit-Cost analysis is covered in Wohl and Martin (1967), Hutchinson (1974), and Stopher and Meybourg (1976). Alternative approaches such as cost-effectiveness analysis are described in Stopher and Meybourg (1976) and Meyer and Miller (1984).

The method proposed for the calculation of surplus variations for transportation system users is original; it extends the "classical" results for aggregate models and those in Williams (1977) for behavioral models. The paper by Jara-Diaz and Friesz (1982) deals with the evaluation of users surplus variations with descriptive demand functions and several dimensions.

Notes

(1) Complementary projects reciprocally increase their positive effects (e.g. park and ride facilities and railways lines), while integrative projects aim at reducing the reciprocal negative effects (e.g. park pricing and upgrading public transport).

(2) This assumes that mathematical models are used to simulate the relevant effects of hypothetical projects exogenously specified. This approach is the most commonly used in applications. However, mathematical models can also been used as supply design tools, as discussed in Chapter 9. As stressed in that chapter supply design models are generally relative to certain types of project (e.g. traffic-signal control or transit lines frequencies) included in wider system projects. In most cases

SDM should be seen as generators of alternative supply states rather than as tools to get the "optimal" solution. For these reasons supply design models can be included, at least conceptually, in the overall system of mathematical models.

(3) Monitoring has a conceptual function analogous to feedback in "closed loop" control systems, which usually prove to be more efficient than open loop systems.

(4) In reality, companies operating transportation services also have several objectives and/or must take into account the impacts of their decisions on different subjects. Economic analysis, in the broad sense, should be extended to all the main decision-makers who operate in a transportation system, though with different objectives and constraints.

(5) As will be seen more clearly in the following, effects for the users are measured as variations induced in their choices.

(6) The "economic life" of a project can be defined conventionally as the period of validity of the project. In the case of infrastructure, this is period for which no major extraordinary maintenance works are necessary. The arbitrariness of this definition is partly compensated for by the possibility of a residual value of the project at the end of the period under consideration.

(7) As stated in Chapter 5, the class is a group of users sharing the same behavioral parameters relevant to the specific application. A user class is usually defined by the pair: socio-economic category, trip purpose. In the limiting case of completely disaggregate models, the class i may coincide with a single individual.

(8) Extra trips undertaken because of the effect of the generalized cost reduction are sometimes indicated as the demand generated by the project P. This term should be better qualified since in some applications trips diverted from other destinations, modes or even paths are referred to as generated demand. To be consistent with the general system approach followed in this book, these trips should be seen as diverted trips or demand, while generated trips are those which wouldn't be made at all in the NP state.

(9) The integral (10.4.15) depends only on the extremes of integration if the Jacobian of demand functions is symmetrical with respect to generalized path costs:

$$\frac{\partial d_1}{\partial g_2} = \frac{\partial d_2}{\partial g_1}$$

This condition is seldom, if ever, met by usual demand models.

(10) In spite of the advantages of the behavioral approach to the calculation of surplus variation, in applications descriptive models are often adopted, even when demand models have Logit or other random utility specifications. This can be explained, at least partly, by the persistence of tradition.

(11) Notice that Expected Maximum Perceived Utility can increase both for a reduction in transportation costs and for an increase in the attractiveness of same zones.

(12) The descriptive approach was introduced to deal with the case in which the project results in the reduction of generalized transportation costs, particularly for road systems. Furthermore, the implicit demand models were often deterministic and this, as it has been shown, implies that there will be no paradoxical results. These conditions, however, are not necessarily met by all the applications of the method.

(13) Notice that this approach is equivalent to the calibration of implicit utility functions of the decision-maker on the basis of revealed and/or stated preferences. It is conceptually analogous to the calibration of demand models with utility functions for transportation related choices and can be solved by using the parameters estimation techniques described in Chapter 8.

(14) For the monotonicity of the transformations used, equation (10.5.5) would hold if l_{mj} or u_{mj} were used instead of x_{mj}.

(15) This would be the case for continuous variables supply design problem discussed in Chapter 9. The main difference with the problems described in Chapter 9, though, is that there are several objective functions (indicators) rather than a single objective function.

A REVIEW OF NUMERICAL ANALYSIS

This appendix contains an overview of the main results of numerical analysis used for the formulation, analysis and solution of the mathematical models described in the text.

A.1. Sets and functions

A.1.1. Elements of set topology

In this section some properties of numerical sets are outlined, with reference to the n-dimensional Euclidean space E^n. Numerical sets are made up of points in E^n, i.e. vectors (assumed to be column vectors) with n real components $\boldsymbol{x}^T = (x_1, \ldots, x_n)$, among which the Euclidean norm (or module) $||\boldsymbol{x}||=(\Sigma_i\, x_i^2)^{1/2} = (\boldsymbol{x}^T\boldsymbol{x})^{1/2}$ and the corresponding Euclidean distance are defined.

The sphere of radius δ and center $\boldsymbol{x}$ is defined a neighborhood, $N_\delta(\boldsymbol{x})$, of radius δ of the point $\boldsymbol{x} \in E^n$:

$$N_\delta(\boldsymbol{x}) = \{y : || y - \boldsymbol{x} || < \delta\}$$

A point $\boldsymbol{x} \in E^n$ is said to be interior to the set $S \subseteq E^n$, if there is at least a *neighborhood of finite radius* δ entirely contained in S. A point $\boldsymbol{x} \in E^n$ is at the boundary of the set S if all the neighborhoods of $\boldsymbol{x}$, however small the radius δ, contain points belonging and points not belonging to S. A nonempty set S is said to be *open* if all the points belonging to S are interior points, i.e. if no boundary point belongs to the set; S is closed if all the boundary points belong to the set. A set S is said to be *limited* if (for all the points belonging to it) a neighborhood of finite radius including all the points of the set can be found:

$$\forall \boldsymbol{x} \in S \quad \exists \delta > 0,\ \delta \text{ finite} : S \subseteq N_\delta(\boldsymbol{x})$$

A closed and limited subset of E^n is *compact* (and vice versa).

For example, the set $S = \{(x_1, x_2) : x^2_1 + x^2_2 \leq 1\}$ of the points belonging to the circle with unitary radius and center in the origin is a closed and limited set, the set

$S_1 = \{(x_1, x_2) : x^2_1 + x^2_2 < 1\}$ also called *intS*, is an open and limited set. The boundary of S and S_1 consists of the set $S_2 = \{(x_1, x_2) : x^2_1 + x^2_2 = 1\}$.

Given two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the set of points $\boldsymbol{x}$ defined by:

$$\{\boldsymbol{x} : \boldsymbol{x} = \mu \boldsymbol{x}_1 + (1 - \mu)\, \boldsymbol{x}_2,\ \mu \in [0,1]\}$$

is called a segment of extremes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

A nonempty set S is said to be *convex* if all the points of the segment joining any two points belonging to the set, belong to the set itself (Fig. A.1.1). In formal terms, it yields:

$$\boldsymbol{x} = \mu \boldsymbol{x}_1 + (1 - \mu)\, \boldsymbol{x}_2 \in S \qquad \forall\, \mu \in [0,1],\ \forall\, \boldsymbol{x}_1, \boldsymbol{x}_2 \in S \qquad \text{(A.1.1)}$$

CONVEX SET NON-CONVEX SET

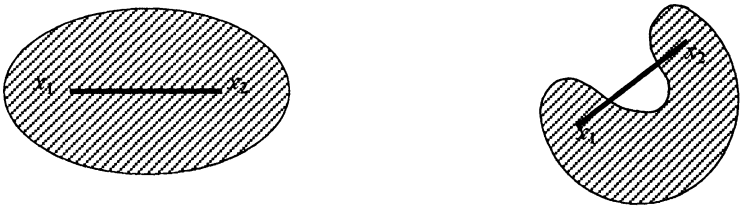


Fig. A.1.1 Illustration of convex and non-convex sets.

The intersection of convex sets is a convex set. Sets defined by a system of linear equalities and/or inequalities, also known as *polyhedral sets*, are convex sets.

In fact, given $S = \{\boldsymbol{x} : \boldsymbol{A}\, \boldsymbol{x} \leq \boldsymbol{b}\}$ if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belong to S, we obtain:

$$\boldsymbol{A}\, [\mu \boldsymbol{x}_1 + (1 - \mu)\, \boldsymbol{x}_2] = \mu \boldsymbol{A}\, \boldsymbol{x}_1 + (1 - \mu)\, \boldsymbol{A}\, \boldsymbol{x}_2 \leq \mu \boldsymbol{b} + (1 - \mu)\, \boldsymbol{b} = \boldsymbol{b} \qquad \mu \in [0,1]$$

Any point $\boldsymbol{x}$ belonging to the segment with extremes $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belongs to S. An analogous demonstration can be repeated for the set $S \equiv \{\boldsymbol{x} : \boldsymbol{A}\, \boldsymbol{x} = \boldsymbol{b}\}$.

Given a point $\boldsymbol{x}^*$, for each non-null vector (direction) $\boldsymbol{h} \neq \boldsymbol{0}$, the set of points lying on the half-line of origin $\boldsymbol{x}^*$ and direction defined by the vector $\boldsymbol{h}$ is *a ray* emanating from $\boldsymbol{x}^*$ along direction $\boldsymbol{h}$. This set is formally defined by:

$$\{\boldsymbol{x} : \boldsymbol{x} = \boldsymbol{x}^* + \mu \boldsymbol{h},\ \ \mu \geq 0\}$$

A vector $\boldsymbol{h}$ is a feasible direction at the point $\boldsymbol{x}^*$ for the set S, if it is possible to move along the direction of a finite quantity from the point $\boldsymbol{x}^*$ and remain within the set S:

$$\exists \mu^* > 0 : \boldsymbol{x} = \boldsymbol{x}^* + \mu \boldsymbol{h} \in S \quad \forall \mu < \mu^* , \mu \geq 0$$

Given a set S, the set $D(\boldsymbol{x}^*)$ of the feasible directions at a point $\boldsymbol{x}^*$ belonging to S (Fig. A.1.2) is formally defined as:

$$D(\boldsymbol{x}^*) = \{\boldsymbol{h} \neq \boldsymbol{0} : \exists \mu^* > 0 : \boldsymbol{x} = \boldsymbol{x}^* + \mu \boldsymbol{h} \in S \quad \forall \mu < \mu^*, \ \mu \geq 0\}$$

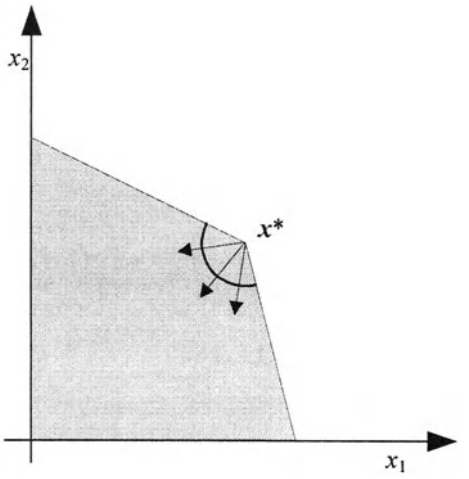


Fig. A.1.2 Illustration of feasible directions.

For a convex set S, the set of feasible directions at point $\boldsymbol{x}^*$ can also be defined as:

$$\{\boldsymbol{h} = (\boldsymbol{x} - \boldsymbol{x}^*) : \boldsymbol{x} \in S\}$$

A.1.2. Differentiable functions

A scalar-valued function of a vector $y = f(\boldsymbol{x})$, with values in E^1 and defined on an open set $S \subseteq E^n$ is said to be continuous at point $\boldsymbol{x}^* \in S$ if small variations of the variables $\boldsymbol{x}$ induce small variations of the variable y. Formally, the function $y = f(\boldsymbol{x})$ is said to be continuous at point $\boldsymbol{x}^*$ if for any neighborhood $N_\delta(y^*)$ of the point $y^* = f(\boldsymbol{x}^*)$, however small, there is a neighborhood $N_\varepsilon(\boldsymbol{x}^*)$ of the point $\boldsymbol{x}^*$ such that the points $\boldsymbol{x}$ belonging to them have values $y = f(\boldsymbol{x})$ in the neighborhood of y^*:

$$\forall \ \delta > 0, \exists \ \varepsilon > 0 : y = f(\boldsymbol{x}) \in N_\delta(y^* = f(\boldsymbol{x}^*)) \quad \forall \boldsymbol{x} \in N_\varepsilon(\boldsymbol{x}^*)$$

A scalar function of vector $f(\boldsymbol{x})$ with values in E^1 and defined on a closed set $S \subseteq E^n$ is said to be differentiable at the point $\boldsymbol{x}^* \in S$ if there is a vector, known as *gradient* of the function in the point and denoted by $\nabla f(\boldsymbol{x}^*)$, such that the

difference between the value of the function at any point $x \in S$ and its linear approximation in x^* along $\nabla f(x^*)$, given by $f(x^*) + \nabla f(x^*)^T (x - x^*)$, is an infinitesimal of superior order with respect to the norm of the vector $(x - x^*)$:

$$\lim_{x \to x^*} \frac{f(x) - f(x^*) - \nabla f(x^*)^T (x - x^*)}{\|x - x^*\|} = 0 \qquad \forall x \in S \qquad \text{(A.1.2)}$$

The components of the vector $\nabla f(x^*)$ are the partial derivatives of the function:

$$\nabla f(x^*)^T = \left[\frac{\partial f(x^*)}{\partial x_1}, \frac{\partial f(x^*)}{\partial x_2}, \ldots, \frac{\partial f(x^*)}{\partial x_n} \right] \qquad \text{(A.1.3)}$$

A function with first continuous partial derivatives is differentiable, and also continuous.

The gradient of a function can be represented in the space E^n with the same dimensionality of the definition set S. In the same space can be defined the level curves of the function (loci of the points x to which the same value of $f(x)$ corresponds). The gradient in each point x^* is a vector perpendicular to the tangent at the level curve $f(x^*)$ and, as will be seen, points towards increasing values of the function (see Fig. A.1.3).

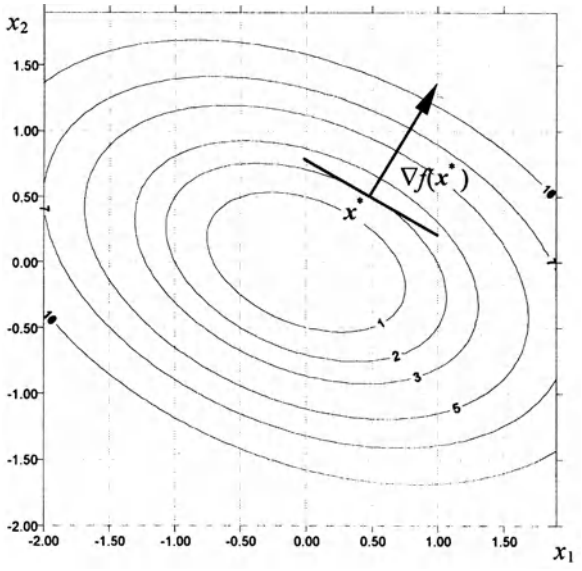


Fig. A.1.3 Level curves and gradient.

Given a scalar function $f(\boldsymbol{x})$, defined in S, a point $\boldsymbol{x}^* \in S$ and a direction vector $\boldsymbol{h}$ such that $\boldsymbol{x}^* + \mu \boldsymbol{h} \in S$ for values of μ less than μ^*, the *directional derivative* of the function in $\boldsymbol{x}^*$ along the direction $\boldsymbol{h}$ can be defined as the limit:

$$f'(\boldsymbol{x}^*, \boldsymbol{h}) = \lim_{\mu \to 0} \frac{f(\boldsymbol{x}^* + \mu \boldsymbol{h}) - f(\boldsymbol{x}^*)}{\mu} \tag{A.1.4}$$

If $f(\boldsymbol{x})$ is differentiable in $\boldsymbol{x}^*$, it is easy to demonstrate that the directional derivative can be expressed in terms of the gradient:

$$f'(\boldsymbol{x}^*, \boldsymbol{h}) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} \tag{A.1.5}$$

A direction $\boldsymbol{h}$ along which it is possible to move by a finite quantity starting from $\boldsymbol{x}^*$, increasing the value of the function, at least in a neighborhood of $\boldsymbol{x}^*$, is known as an ascent direction. In other words, a direction $\boldsymbol{h}$ is an ascent direction if a positive scalar θ^* can be found such that for each $0 < \theta < \theta^*$ it results:

$$f(\boldsymbol{x}^* + \theta \boldsymbol{h}) > f(\boldsymbol{x}^*) \tag{A.1.6}$$

It can be demonstrated (by using the property of the directional derivative (A.1.5) and the theorem of sign permanence) that a direction $\boldsymbol{h}$ is an ascent direction if and only if the directional derivative of $f(\boldsymbol{x})$ at the point $\boldsymbol{x}^*$ along the direction $\boldsymbol{h}$ is positive:

$$f'(\boldsymbol{x}^*, \boldsymbol{h}) = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} > 0 \tag{A.1.7}$$

Similarly, the directions along which it is possible to move starting from $\boldsymbol{x}^*$, causing a decrease of the function value are known as descent directions, and have negative directional derivative at $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} < 0$.

The gradient of a differentiable function, at whatever point it is different from zero, is an ascent direction. In fact, under the assumptions made, it results:

$$f'(\boldsymbol{x}^*, \nabla f(\boldsymbol{x}^*)) = \nabla f(\boldsymbol{x}^*)^T \nabla f(\boldsymbol{x}^*) = ||\nabla f(\boldsymbol{x}^*)||^2 > 0 \tag{A.1.8}$$

Vice versa, the direction opposite to the gradient $-\nabla f(\boldsymbol{x}^*)$, if different from zero, is a descent direction of $f(\boldsymbol{x})$ in $\boldsymbol{x}^*$.

A scalar function $f(\boldsymbol{x})$ is said to be *doubly* or *twice differentiable* in $\boldsymbol{x}$ if there is a vector $\nabla f(\boldsymbol{x}^*)$ and a symmetrical matrix $\boldsymbol{H}_f(\boldsymbol{x}^*)$ of dimensions $(n \times n)$ such that:

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}^*} \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) - 1/2 (\boldsymbol{x} - \boldsymbol{x}^*)^T \boldsymbol{H}_f(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)}{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2} = 0 \quad \forall \boldsymbol{x} \in S \tag{A.1.9}$$

Equation (A.1.9) expresses the condition that the difference between the value of the function and its quadratic approximation is an infinitesimal of superior order with respect to the square norm of the vector $(\boldsymbol{x} - \boldsymbol{x}^*)$. The matrix $\boldsymbol{H}_f(\boldsymbol{x}^*)$ is called the *Hessian matrix* of $f(\boldsymbol{x})$ at $\boldsymbol{x}^*$ and its components are the second order partial derivatives of $f(\boldsymbol{x})$ at $\boldsymbol{x}^*$:

$$\boldsymbol{H}_f(\boldsymbol{x}^*) = \begin{vmatrix} \dfrac{\partial^2 f(\boldsymbol{x}^*)}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f(\boldsymbol{x}^*)}{\partial x_1 \partial x_n} \\ \cdots & & \cdots \\ \cdots & & \cdots \\ \cdots & & \cdots \\ \dfrac{\partial^2 f(\boldsymbol{x}^*)}{\partial x_1 \partial x_n} & \cdots & \dfrac{\partial^2 f(\boldsymbol{x}^*)}{\partial x_n^2} \end{vmatrix} \qquad \text{(A.1.10)}$$

A function is *doubly differentiable* if it has continuous second partial derivatives. In this case the first partial derivatives are differentiable (because they have continuous partial derivatives), the function is differentiable and therefore continuous. Furthermore the second partial derivatives are not dependent on the order of derivation and the Hessian matrix is symmetric.

Taylor's formulae of the first and second order relative to the scalar function $f(\boldsymbol{x})$ around the point $\boldsymbol{x}^*$ are respectively:

$$\exists\, \boldsymbol{x}^\circ \in (\boldsymbol{x}^*, \boldsymbol{x}) : f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^\circ)^T (\boldsymbol{x} - \boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in S \qquad \text{(A.1.11)}$$

$$\exists\, \boldsymbol{x}^\circ \in (\boldsymbol{x}^*, \boldsymbol{x}) : f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) + 1/2 (\boldsymbol{x} - \boldsymbol{x}^*)^T \boldsymbol{H}_f(\boldsymbol{x}^\circ) (\boldsymbol{x} - \boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in S \qquad \text{(A.1.12)}$$

where $\boldsymbol{x}^\circ$ is a point within to the segment $(\boldsymbol{x} - \boldsymbol{x}^*)$.

Equations (A.1.11) and (A.1.12) obviously require $f(\boldsymbol{x})$ to be differentiable of the first and second order respectively.

An m-vectorial function $\boldsymbol{g}(\boldsymbol{x})$ associates a vector of m components, i.e. a point of E^m, to an n dimensional vector, i.e. a point of E^n; i.e., it is a vector in m functions:

$$\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x})]^T$$

which associates to each n-dimensional vector $\boldsymbol{x} \in S$ an m-dimensional vector $\boldsymbol{g}(\boldsymbol{x})$. The function $\boldsymbol{g}(\boldsymbol{x})$ is said to be differentiable at the point $\boldsymbol{x}^*$ if all its component functions are differentiable. The *Jacobian matrix* of $\boldsymbol{g}(\boldsymbol{x})$ is a matrix of dimensions $(m \times n)$ which has the gradients of the component functions $g_i(\boldsymbol{x})$ as its rows:

$$\boldsymbol{Jac}[\boldsymbol{g}(\boldsymbol{x}^*)] = \begin{vmatrix} \dfrac{\partial g_1(\boldsymbol{x}^*)}{\partial x_1} & \cdots & \dfrac{\partial g_1(\boldsymbol{x}^*)}{\partial x_n} \\ \cdots & & \cdots \\ \cdots & & \cdots \\ \cdots & & \cdots \\ \dfrac{\partial g_m(\boldsymbol{x}^*)}{\partial x_1} & \cdots & \dfrac{\partial g_m(\boldsymbol{x}^*)}{\partial x_n} \end{vmatrix} \tag{A.1.13}$$

An n-vectorial function of vector $\boldsymbol{g}(\boldsymbol{x})$ (in this case $m = n$) defined in a set $S \subseteq E^n$ is strictly increasing monotone if for each pair of different points $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in S$ it results:

$$(\boldsymbol{g}(\boldsymbol{x}_1) - \boldsymbol{g}(\boldsymbol{x}_2))^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) > 0 \quad \forall \boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in S \tag{A.1.14}$$

The function is said to be non-decreasing monotone if weak inequality holds (≥ 0). Similarly, functions can be denoted as strictly decreasing or non-increasing monotone if the reversed inequalities hold. If the two points $\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in S$, differ only in the i-th component, i.e. $x_{1,i} \neq x_{2,i}$, with $x_{1,j} = x_{2,j} \ \forall j \neq i$, from inequality (A.1.14) it follows:

$$(g_i(x_{1,i}) - g_i(x_{2,i}))^T (x_{1,i} - x_{2,i}) > 0$$

hence all the component functions are increasing monotone functions of every component of the vector $\boldsymbol{x}$ for given values of all other components (scalar functions of scalar).

If the Jacobian matrix $\boldsymbol{Jac}[\boldsymbol{g}(\boldsymbol{x})]$ of the function $\boldsymbol{g}(\boldsymbol{x})$, assumed to be differentiable, is positive (negative) semi-definite over the whole set of definition S, the function $\boldsymbol{g}(\boldsymbol{x})$ is non-decreasing (non-increasing monotone). If the Jacobian is positive (negative) definite, the function is monotone strictly increasing (decreasing).

A.1.3. Convex functions

A scalar function of vector $f(\boldsymbol{x})$ defined in the convex set $S \subseteq E^n$ is denoted convex if for any pair of points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belonging to S the following relationship holds:

$$f[\mu \boldsymbol{x}_1 + (1 - \mu)\boldsymbol{x}_2] \leq \mu f(\boldsymbol{x}_1) + (1 - \mu) f(\boldsymbol{x}_2) \quad \forall \mu \in [0,1] \tag{A.1.15}$$

The geometrical interpretation of (A.1.15) is that the value of the function calculated at whatever point of the segment joining $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is not greater than the linear combination of the values calculated at the endpoints, see Fig. A.1.4.

It can be demonstrated that a differentiable function is convex if and only if it satisfies the following condition:

$$f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}_1) + \nabla f(\boldsymbol{x}_1)^T (\boldsymbol{x}_2 - \boldsymbol{x}_1) \quad \forall \boldsymbol{x}_2, \boldsymbol{x}_1 \in S \subseteq E^n \tag{A.1.16}$$

i.e., if the value of the function in $\boldsymbol{x}_2$ is not lower than the value of its linear extrapolation starting from $\boldsymbol{x}$ see Fig. A.1.5. By inverting points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, in (A.1.16) and summing the two expressions we also get:

$$(\nabla f(\boldsymbol{x}_1) - \nabla f(\boldsymbol{x}_2))^T (\boldsymbol{x}_1 - \boldsymbol{x}_2) \geq 0 \tag{A.1.17}$$

i.e., the gradient of a convex differentiable function is a monotone non-decreasing vectorial function of the vector $\boldsymbol{x}$.

It can also be shown that the necessary and sufficient condition that a doubly differentiable function is convex is that its Hessian matrix is positive semi-definite over the whole set of definition S:

$$\boldsymbol{x}^T \boldsymbol{H}(\boldsymbol{x}^*) \boldsymbol{x} \leq 0 \quad \forall \boldsymbol{x}, \boldsymbol{x}^* \in S \tag{A.1.18}$$

If the inequalities (A.1.15), (A.1.16) and (A.1.17) and (A.1.18) hold with the sign of strict inequality, the function is said to be strictly convex.

A function $f(\boldsymbol{x})$ given by a linear combination with positive coefficients of convex functions $f^i(\boldsymbol{x})$ is convex:

$$f(\boldsymbol{x}) = \Sigma_i \mu_i f^i(\boldsymbol{x}) \qquad \mu_i > 0$$

It is also strictly convex if at least one of the component functions is strictly convex.

The function $f(\boldsymbol{x})$ is said to be (strictly) concave if $-f(\boldsymbol{x})$ is (strictly) convex. In this case $f(\boldsymbol{x})$ verifies (A.1.15), (A.1.16), (A.1.17) and (A.1.18) with the inequalities inverted. A linear function is both convex and concave since (A.1.15) holds with the sign of equality.

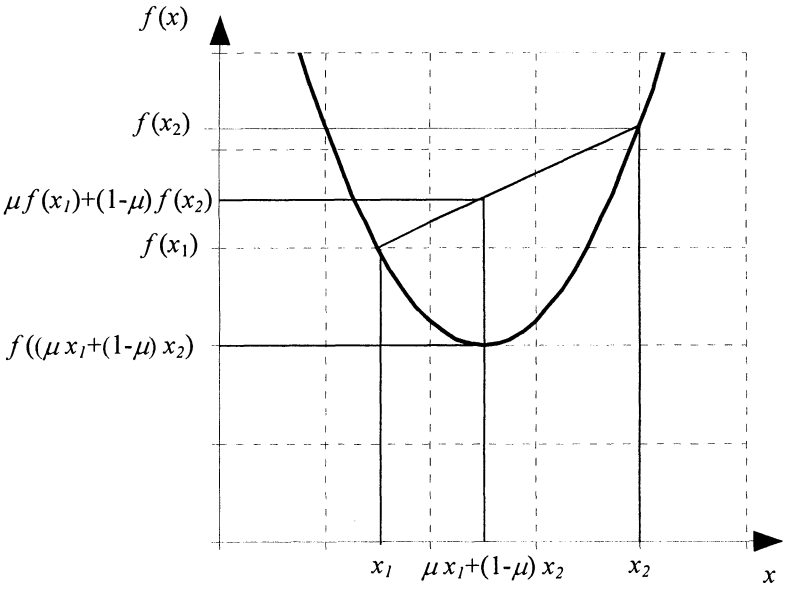


Fig. A.1.4 Geometrical interpretation of the definition of convex function.

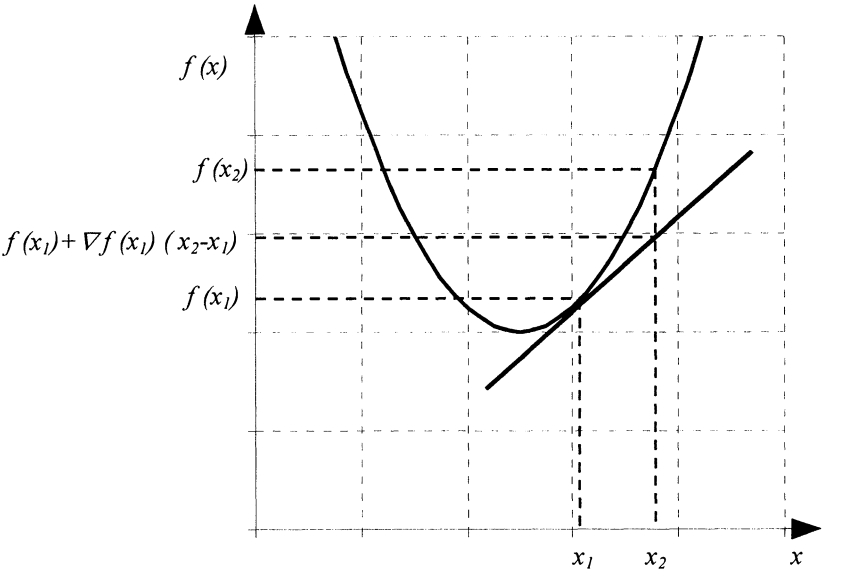


Fig. A.1.5 Geometrical interpretation of the convexity of a differentiable function.

A.2. Solution algorithms

A mathematical problem with a solution given by a vector $\boldsymbol{x}^* \in S$ is said solvable in closed form if there is a relationship allowing the calculation of the solution (or solutions) of the problem as a function of the parameters of the problem itself.

Consider, for example, the problem of searching for the null points of a function $f(\boldsymbol{x})$ in the sets, i.e. the problem of solving the equation $f(\boldsymbol{x}) = 0$, with the condition that the solutions belong to the set S. In the case of a second-order polynomial function, $a\,x^2 + b\,x + c$, the null points are a solution to the equation: $a\,x^2 + b\,x + c = 0$. As is known, the equation has two solutions x_1 and x_2 (real or conjugate complex) which can be calculated in closed form by means of the formula: $x_{1,2} = (-b \pm (b^2 - 4\,a\,c)^{1/2})/(2\,a)$.

More in general, when a closed-form solution cannot be found, recursive equations generating a succession of points $\{\boldsymbol{x}^1,..., \boldsymbol{x}^k, \boldsymbol{x}^{k+1},...\}$ are adopted, i.e:

$$\boldsymbol{x}^{k+1} = \varphi(\boldsymbol{x}^k) \tag{A.2.1}$$

The equation (A.2.1) defines an algorithm solving the problem, if the recursive equation stops in the solution being sought $\boldsymbol{x}^*$:

$$\boldsymbol{x}^* = \varphi(\boldsymbol{x}^*)$$

and vice versa if it is found that the point at which the equation stops $\boldsymbol{x}^* = \varphi(\boldsymbol{x}^*)$, is the solution sought.

An algorithm is said to be feasible if all the elements of the succession belong to the set of feasible solutions, $\boldsymbol{x}^k \in S$. An algorithm is convergent in a finite number of steps if it can be demonstrated that there is a finite number n such that $\boldsymbol{x}^n = \boldsymbol{x}^*$; a closed form solution therefore is an algorithm convergent in one step. A resolutive algorithm is said to be asymptotically convergent if it can be demonstrated that the succession of points converges to the solution sought: i.e., $lim_{k\to\infty}\, \boldsymbol{x}^k = \boldsymbol{x}^*$. If no form of convergence can be demonstrated, the algorithm is said to be heuristic.

A.3. Fixed point problems

Let $\psi(\boldsymbol{x})$ be a n-vectorial function of a vector $\boldsymbol{x}$ defined in a set $S \subseteq E^n$, with values in the set $T = \psi(S) = \{\psi(\boldsymbol{x}) : \boldsymbol{x} \in S\} \subseteq E^n$; the point $\boldsymbol{x}^* \in S$ is denoted fixed point if the function has a value equal to the argument (see Fig. A.3.1):

$$\boldsymbol{x}^* = \psi(\boldsymbol{x}^*) \qquad \boldsymbol{x}^* \in S \tag{A.3.1}$$

Note that specifying a solution algorithm for any mathematical problem based on the recursive equation (A.2.1) is equivalent to defining a function $\varphi(x)$ having as its fixed point the solution of the mathematical problem under study.

Fixed point problems, found in various branches of engineering and economics, can easily be related to nonlinear systems of equations (and vice versa):

$$x^* - \psi(x^*) = 0 \qquad x^* \in S$$

A particularly interesting case of fixed point problem, called the *compound fixed point* problem, is identified in the search for equilibrium configurations between two vectors, $x \in S_x \subseteq E^n$ and $y \in S_y \subseteq E^m$ (also with $n \neq m$) which reciprocally influence each other (see Fig. A.3.1b), i.e:

$$\begin{cases} y^* = \eta(x^*) & x^* \in S_x \quad y^* \in S_y \\ x^* = \rho(y^*) & y^* \in S_y \quad x^* \in S_x \end{cases} \tag{A.3.2}$$

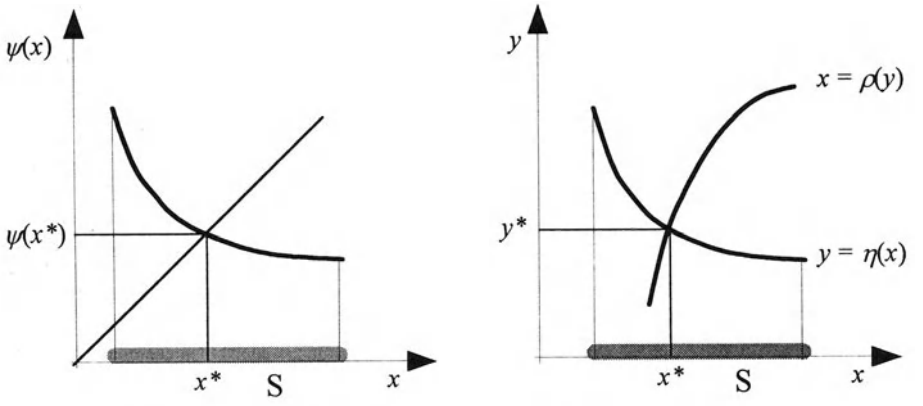


Fig. A.3.1 Simple and compound fixed points.

In fact, by combining the previous relationships, a compound fixed point problem in the variable x is obtained:

$$x^* = \rho(\eta(x^*)) \qquad x^* \in S_x \tag{A.3.3a}$$

with $\eta(S_x) \subseteq S_y$ and $\rho(\eta(S_x)) \subseteq S_x$. Similarly, an equivalent[(1)] fixed point problem in the variable y can be defined:

$$y^* = \eta(\rho(y^*)) \qquad y^* \in S_y \tag{A.3.3b}$$

with $\rho(S_y) \subseteq S_x$ and $\eta(\rho(S_y)) \subseteq S_y$.

The properties of the nonlinear equations system (A.3.2) or of each of the two compound fixed point problems (A.3.3a-b) depend on the characteristics of the two functions involved, $y = \eta(x)$ and $x = \rho(y)$, and on the sets of definition of the variables, S_x and S_y.

A.3.1. Properties of fixed points

Sufficient conditions for the existence and uniqueness of the solution of a fixed point problem are given by Banach's theorem[2] which also allows the specification of an asymptotically convergent resolutive algorithm. Only a restricted class of functions, however, satisfies these conditions. What follows, therefore, describes some of the weaker conditions (some of these conditions can be extended with some mathematical complications). Sufficient conditions for the existence of at least one solution of the fixed point problem (A.3.1), i.e. for the existence of at least one fixed point of a function, are given by Brouwer's theorem stated below.

Brouwer's theorem. The fixed point problem (A.3.1) has at least one solution, i.e. the function $\psi(x)$ defined in the set $S \subseteq E^n$ with values in the set $T = \psi(S) \subseteq E^n$ has at least one fixed point if:

T is a subset of S, $T \subseteq S$, i.e. $\psi(x) \in S \quad \forall x \in S$;
S is a compact and convex nonempty set;
$\psi(x)$ is a continuous function.

The application of Brouwer's theorem to compound fixed point problems, such as the one defined by (A.3.3.a), requires both the functions $\eta(x)$ and $\rho(y)$ to be continuous, the definition set to be a nonempty, compact and convex set, and $S_x \subseteq \rho(\eta(S_x))$, i.e. $\rho(\eta(x)) \in S_x \quad \forall x \in S_x$.

A graphic illustration of the relevance of some of the assumptions of Brouwer's theorem is given in Fig. A.3.2.

Sufficient conditions for the uniqueness of the solution of the fixed point problem (A.3.1), i.e. for the existence of at most one fixed point of a function are given by the simple theorem described below.

Theorem. The fixed point problem (A.3.1) has at most one solution, i.e. the function defined in the set $\psi(x)$ with values in the set $S \subseteq E^n$ has at most one fixed point, if $T = \psi(x) \subseteq E^n$.

$\psi(x)$ is a monotone non-increasing[3] function, over the whole set S:

$$(\psi(x') - \psi(x''))^T (x' - x'') \leq 0 \qquad \forall x', x'' \in S$$

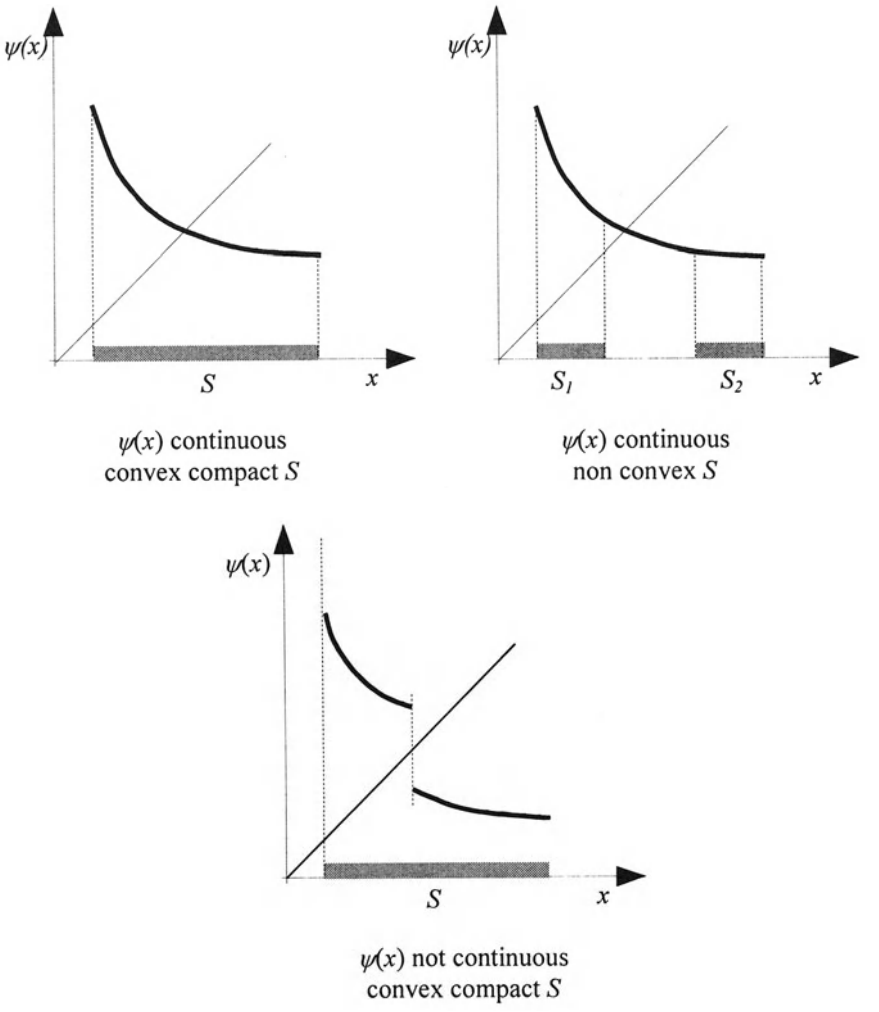


Fig. A.3.2 Illustration of the assumptions of Brouwer's theorem.

In fact, if there existed two different fixed point vectors, $x_1^* \neq x_2^* \in S$, being $x_1^* = \psi(x_1^*)$, for the monotonicity of the function $\psi(x)$ it would follow:

$$||(x_1^* - x_2^*)||^2 = (x_1^* - x_2^*)^T (x_1^* - x_2^*) \leq 0$$

In contradiction to the condition $||(x_1^* - x_2^*)||^2 > 0$ for any $x_1^* \neq x_2^*$.

Uniqueness conditions can be extended to compound fixed point problems, such as that defined by (A.3.3.a), in the case of two monotone functions in the opposite direction (and at least one of the two strictly monotone), as shown in the theorem described below.

Theorem. The compound fixed point problem (A.3.3a) has at the most one solution, i.e. the compound function $\psi(\boldsymbol{x}) = \rho(\eta(\boldsymbol{x}))$ defined in the set $S_x \subseteq E^n$ with $\eta(S_x) \subseteq S_y$ and $\rho(\eta(S_x)) \subseteq S_x$, has at most one fixed point if the two functions $\rho(.)$ and $\eta(.)$ are monotone in the opposite direction. For example:

$\boldsymbol{y} = \eta(\boldsymbol{x})$ is a *strictly increasing function*, i.e:

$$(\eta(\boldsymbol{x}') - \eta(\boldsymbol{x}''))^T (\boldsymbol{x}' - \boldsymbol{x}'') > 0 \qquad \forall \boldsymbol{x}' \neq \boldsymbol{x}'' \in S_x$$

$\boldsymbol{x} = \rho(\boldsymbol{y})$ is a *non-increasing function*, i.e.:

$$(\rho(\boldsymbol{y}') - \rho(\boldsymbol{y}''))^T (\boldsymbol{y}' - \boldsymbol{y}'') \leq 0 \qquad \forall \boldsymbol{y}', \boldsymbol{y}'' \in \eta(S_x)$$

In fact, if there existed two different fixed point vectors, $\boldsymbol{x}_1^* \neq \boldsymbol{x}_2^* \in S$, i.e. $\boldsymbol{x}_1^* = \rho(\eta(\boldsymbol{x}_1^*))$ and $\boldsymbol{x}_2^* = \rho(\eta(\boldsymbol{x}_2^*))$, denoted $\boldsymbol{y}_1^* = (\eta(\boldsymbol{x}_1^*)$ and $\boldsymbol{y}_2^* = \rho(\boldsymbol{x}_2^*)$, from which $\boldsymbol{x}_1^* = \rho(\boldsymbol{y}_1^*)$ and $\boldsymbol{x}_2^* = \rho(\boldsymbol{y}_2^*)$, for the monotonicity of the function $\rho(\boldsymbol{y})$ it would follow:

$$(\boldsymbol{x}_1^* - \boldsymbol{x}_2^*)^T (\boldsymbol{y}_1^* - \boldsymbol{y}_2^*) = (\rho(\boldsymbol{y}_1^*) - \rho(\boldsymbol{y}_2^*))^T (\boldsymbol{y}_1^* - \boldsymbol{y}_2^*) \leq 0$$

In contradiction to the monotonicity of the function $\eta(\boldsymbol{x})$, for $\boldsymbol{x}_1^* \neq \boldsymbol{x}_2^*$:

$$(\boldsymbol{y}_1^* - \boldsymbol{y}_2^*)^T (\boldsymbol{x}_1^* - \boldsymbol{x}_2^*) = (\eta(\boldsymbol{x}_1^*) - \eta(\boldsymbol{x}_2^*))^T (\boldsymbol{x}_1^* - \boldsymbol{x}_2^*) > 0$$

A.3.2. Solution algorithms for fixed point problems

In general, solution algorithms for solving fixed point problems are more recent and less developed than those for optimization problems, to be described in the next section.

Algorithms for fixed point problems (A.3.1) are usually based on the explicit calculation of the Jacobian of the function $\psi(\boldsymbol{x})$, and eventually on the calculation of its eigenvalues, or an estimate of them. This approach is generally difficult to apply to large-scale problems; for this reason what follows will describe some solution algorithms whose application requires only the calculation of the function $\psi(\boldsymbol{x})$. In particular, given a sequence $\{\mu_k\}_{k>0}$ satisfying the condition:

$$\sum_{k>0} \mu_k = \infty, \quad \sum_{k>0} \mu^2{}_k < \infty \tag{A.3.4}$$

an algorithm for the solution of a fixed point problem can be specified by the following recursive equation:

$$x^{k+1} = x^k + \mu_k[\psi(x^k) - x^k] \quad \text{i.e. } x^{k+1} = (1 - \mu_k)x^k + \mu_k\psi(x^k) \tag{A.3.5}$$

with $x^1 \in S$.

By using Blum's theorem (not reported here because of its complexity), it can be demonstrated that if the function $\psi(x)$ has a unique fixed point $x^* = \eta(x^*)$, the relationship (A.3.5) defines a sequence convergent[4] to the fixed point x^*, i.e. $lim_{k\to\infty} x^k = x^*$, if the function $\psi(x)$ is continuous and monotone non-increasing and the set S is nonempty, compact and convex (as required by the sufficient conditions of existence and uniqueness). From a practical point of view, the algorithm is stopped when $x^k \cong \psi(x^k)$, e.g. when a norm value of the vector of components $(x^k_i - \psi_i(x^k)) / x^k_i$ is lower than an pre-assigned threshold. Stop tests based on the distance between values of the vector x, between successive iterations, i.e. $x^{k+1} \cong x^k$, are to be avoided since this difference tends to zero because of the structure of the algorithm, regardless of the proximity to the solution of the fixed point problem.

If the sequence $\{\mu_k\}_{k>0}$ also satisfies the condition

$$\mu_k \in (0,1) \tag{A.3.6}$$

the elements of the sequence generated by the relationship (A.3.5) belong to the set S, $x^k \in S$, S being convex. This property is especially useful from a practical point of view because it provides a feasible solution of the problem at whatever iteration the algorithm stops.

The sequence with the largest elements satisfying both the conditions (A.3.4) and (A.3.6) is given by $\{\mu_k = 1/k\}_{k>0}$. In this case the relationship (A.3.5) leads to the so-called *Method of Successive Averages* or *MSA*:

$$x^{k+1} = x^k + (1/k)\,[\psi(x^k) - x^k] \in S$$

i.e. (A.3.7)

$$x^{k+1} = ((k-1)\,x^k + \psi(x^k))/k$$

with $x^1 \in S$.

The above observations can also be applied to the compound fixed point problem (A.3.3a). In this case the relationship (A.3.5) becomes:

$$x^{k+1} = x^k + \mu_k\,[\rho(\eta(x^k)) - x^k] \tag{A.3.8}$$

with $x^1 \in S_x$.

By using Blum's theorem, it can be demonstrated that if the compound function $\rho(\eta(\boldsymbol{x}))$ has a unique fixed point $\boldsymbol{x}^* = \rho(\eta(\boldsymbol{x}^*))$, the relationship (A.3.8) defines a sequence convergent to the fixed point $\boldsymbol{x}^*$, if the function $\boldsymbol{x} = \rho(\boldsymbol{y})$ is continuous and monotone non-increasing, the function $\boldsymbol{y} = \eta(\boldsymbol{x})$ is continuous and strictly monotone increasing, the set S is nonempty, compact and convex (as required by the sufficient conditions of existence and uniqueness), and if the function $\boldsymbol{y} = \eta(\boldsymbol{x})$ has a symmetrical and continuous Jacobian.

A.4. Optimization problems

Optimal points $\boldsymbol{x}^*$ of a scalar function of a vector, $f(\boldsymbol{x})$, are the points corresponding to minimum or maximum values of the function. For simplicity, what follows will make reference only to minimum points[5]. Formally, let $f(\boldsymbol{x})$ be a scalar function defined in a set $S \subseteq E^n$; the point $\boldsymbol{x}^*$ is called a local minimum point of the function if there is a neighborhood $N_\delta(\boldsymbol{x}^*)$ of radius δ such that the following condition holds:

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) \quad \forall\, \boldsymbol{x} \neq \boldsymbol{x}^*, \boldsymbol{x} \in N_\delta(\boldsymbol{x}^*)$$

If this condition holds for all the points of S, the point $\boldsymbol{x}^*$ is called a global minimum point of the function $f(\boldsymbol{x})$ over S. In general, a continuous function $f(\boldsymbol{x})$ over a compact set S always has at least one global minimum point. A function with a unique minimum point is defined unimodal; an example of this kind of function is given by the strictly convex functions defined previously.
The problem of the search for the minimum points $\boldsymbol{x}^*$ of a function is defined minimum or minimization problem, $f(\boldsymbol{x})$ is defined objective function, S feasibility set. The minimization problem is formally expressed as:

$$\boldsymbol{x}^* = \operatorname{argmin} f(\boldsymbol{x}) \qquad \text{(A.4.1)}$$

$$\boldsymbol{x} \in S$$

Minimization problems and fixed point problems are related to each other. In fact the fixed point problem (A.3.1) defined by the function $\psi(\boldsymbol{x})$ is equivalent to a minimum problem defined by the objective function with non-negative values $f(\boldsymbol{x}) = (\psi(\boldsymbol{x}) - \boldsymbol{x})^T(\psi(\boldsymbol{x}) - \boldsymbol{x})$ being $f(\boldsymbol{x}) = 0$ if and only if $\psi(\boldsymbol{x}) - \boldsymbol{x} = 0$.

The definition of local and global minimum points cannot be used in the search for such points since it would require the calculation of $f(\boldsymbol{x})$ over all the points in S and comparison of their values. It is therefore essential to find necessary and/or sufficient conditions for the minimum points expressed in terms of "local" properties of the function. The necessary and sufficient conditions will be reported in the following by differentiating the case in which the minimum point is interior to an open set and that in which it may be on the boundary of a closed set.

A.4.1. Properties of minimum points

A.4.1.1. Properties of minimum points on open sets

The necessary condition for which the point $\boldsymbol{x}^*$ is the local minimum for the differentiable function $f(\boldsymbol{x})$ defined in an open set S is that it is a point of stationariety of the objective function, i.e. is that in $\boldsymbol{x}^*$ we have $\nabla f(\boldsymbol{x}^*) = 0$.

In fact, if $\boldsymbol{x}^*$ is a point interior to S, any direction is feasible. Furthermore, since $\boldsymbol{x}^*$ is a local minimum point, the directional derivative calculated in $\boldsymbol{x}^*$ must be non-negative for any direction:

$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} \geq 0 \qquad \forall \boldsymbol{h} \tag{A.4.2}$$

Since $-\nabla f(\boldsymbol{x}^*)$ is a feasible direction, the condition (A.4.2) holds only if the gradient is null. Note that the nullity of the gradient in $\boldsymbol{x}^*$ is only a necessary condition for $\boldsymbol{x}^*$ to be a local minimum point. In particular, local maximum points also satisfy the same condition.

If $\boldsymbol{x}^*$ is a point of stationariety of the continuous and second order differentiable function $f(\boldsymbol{x})$, with continuous first and second derivatives, the sufficient condition for $\boldsymbol{x}^*$ to be a local minimum is that the Hessian matrix in $\boldsymbol{x}^*$ is positive semi-definite.

In fact, applying Taylor's second-order formula (A.1.12), it follows:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + 1/2(\boldsymbol{x} - \boldsymbol{x}^*)^T \boldsymbol{H}_f(\boldsymbol{x}^\circ)(\boldsymbol{x} - \boldsymbol{x}^*) \tag{A.4.3}$$

where $\boldsymbol{x}^\circ$ is a point of the segment $(\boldsymbol{x}, \boldsymbol{x}^*)$.

If $\boldsymbol{H}_f(\boldsymbol{x})$ is positive definite for the sign permanence theorem, a neighborhood of $\boldsymbol{x}^*$, $N_\delta(\boldsymbol{x}^*)$, can be found such that $\boldsymbol{H}_f(\boldsymbol{x})$ is positive definite at all points within this neighborhood. If $\boldsymbol{x}$ belongs to this neighborhood, all the points of the segment $(\boldsymbol{x}, \boldsymbol{x}^*)$ belong to it and so does $\boldsymbol{x}^\circ$. From this it follows that:

$$1/2\,(\boldsymbol{x} - \boldsymbol{x}^*)^T \boldsymbol{H}_f(\boldsymbol{x}^\circ)(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \Rightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in N_\delta(\boldsymbol{x}) \tag{A.4.4}$$

If the function $f(\boldsymbol{x})$ is convex, the Hessian matrix is positive semi-definite in all S and from (A.4.4) it follows that the nullity of the gradient is a necessary and sufficient condition for $\boldsymbol{x}^*$ being a global minimum point. The minimum points of a convex function make up a convex set. Furthermore, if the function is strictly convex, a point of stationariety is also the unique global minimum point.

A.4.1.2. Properties of minimum points on closed sets

In general, the closed set S is defined by equality and/or inequality relationships, known as *constraints*. In what follows, the case of m inequalities constraints will be discussed, equality constraints can be reduced to two inequalities ($g_i(\boldsymbol{r})=0$ is equivalent to $g_i(\boldsymbol{x}) \leq 0$ and $-g_i(\boldsymbol{x}) \leq 0$):

$$S \equiv \{\boldsymbol{x} : g_i(\boldsymbol{x}) \leq 0 \quad i = 1,2,...,m\}$$

Using the m-vectorial function of vector $g(\boldsymbol{x})$, the constraints can be expressed as:

$$\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$$

With this notation the optimization problem can formally be expressed as:

$$min\, f(\boldsymbol{x})$$

$$\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$$

Unlike the previous case, the minimum point might lie on the boundary of the set S. In this case, not all directions are feasible; in particular, the gradient may not be a feasible direction and the stationariety of the function in $\boldsymbol{x}^*$ hasn't to be verified by a minimum point.

Denoting $D(\boldsymbol{x}^*)$, the set of feasible directions at the minimum point $\boldsymbol{x}^*$, because of the results (A.1.7) and (A.1.8) on directional derivatives, the function must have non-negative directional derivatives in $\boldsymbol{x}^*$ for all the feasible directions:

$$\nabla f(\boldsymbol{x}^*)^T \boldsymbol{h} \geq 0 \quad \forall \boldsymbol{h} \in D(\boldsymbol{x}^*) \tag{A.4.5a}$$

If the set S is convex and $\boldsymbol{x}$ is a point belonging to S, the direction $(\boldsymbol{x} - \boldsymbol{x}^*)$ is feasible by definition and (A.4.5a) becomes:

$$\nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \quad \forall \boldsymbol{x} \in S \tag{A.4.5b}$$

The points satisfying (A.4.5a) or (A.4.5b) are also denoted as virtual minimum points since in general the two conditions are only necessary for the point $\boldsymbol{x}^*$ to be the minimum. They are also sufficient if the objective function $f(\boldsymbol{x})$ is convex. Also in this case the minimum points of a convex function make up a convex set. In the case of a strictly convex function, there is a unique minimum point.

A.4.2. Solution algorithms for optimization problems

This section will describe some solution algorithms for particular optimization problems, which have been mentioned in previous chapters.

A.4.2.1. Mono-dimensional optimization algorithms

These algorithms solve the problem of finding the minimum of a function $f(\theta)$ of a scalar variable θ. If the value θ of minimizing $f(\theta)$in the interval $(\theta_{min}, \theta_{max})$ is indicated with θ^*, the mono-dimensional optimization problem can be expressed

as follows:

$$\theta^* = \underset{\theta_{\min} \le \theta \le \theta_{\max}}{argmin} \; f(\theta) \tag{A.4.6}$$

In practice, the problem (A.4.6) is rarely solved as such. However, it is an element common to many solution algorithms for more complex problems since as will be seen, it allows to obtain the minimum of a vector function along a direction $\boldsymbol{h}$ starting from a point $\boldsymbol{x}^*$. In this case, in fact, the points of the straight line passing from $\boldsymbol{x}^*$ oriented as the vector $\boldsymbol{h}$ are expressed as $\boldsymbol{x}^* + \theta \boldsymbol{h}$ and as θ varies, the points of the whole straight-line ($-\infty < \theta < +\infty$) or of the half-line concordant with $\boldsymbol{h}$ ($\theta > 0$) are described (see section A.1.1).

The most straightforward algorithm solving the problem (A.4.6) is the "uniform search". The interval θ_{min}, θ_{max}, is subdivided into subintervals of equal widths δ with extremes at the "grid points" $\theta_1 = \theta_{min}, \theta_2, \ldots, \theta_n = \theta_{max}$; the objective function is evaluated in each of the n points θ_k and θ^* is the point corresponding to the lower value of the function (Fig. A.4.1). If the function is convex, the actual minimum point is included in the interval $\theta^* \pm \delta$.

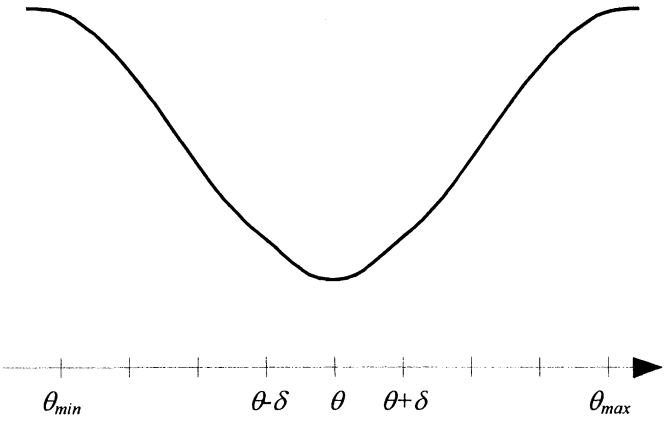


Fig. A.4.1 Uniform search algorithm.

More efficient algorithms for convex functions are based on the principle of "reduction of the uncertainty interval". At each iteration, an interval of extremes (a_k, b_k) is obtained which includes the minimum of the function, called the interval of uncertainty. The width of this interval is reduced at each iteration. In the following, the main steps of one such algorithms, known as the "method of the golden section" will be described. The name derives from its use of the property of the golden section of a segment to re-compute the value of the $f(\theta)$ only once at each iteration see Fig. A.4.2. The algorithm is asymptotically convergent if $f(.)$ is a convex function (even in a weaker sense than that described in section A.1.3)

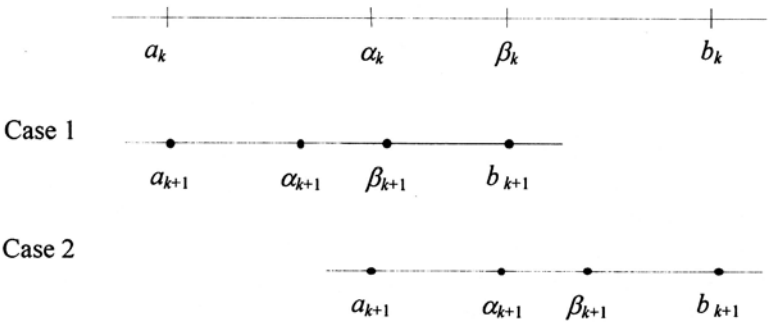


Fig. A.4.2 Illustration of the golden section algorithm.

Golden section algorithm

Step 0 Initialization. The maximum width ε, allowed for the uncertainty interval is chosen. The extremes of the initial interval are set

$$a_1 = \theta_{min}\ b_1 = \theta_{max}$$

This gives the points:

$$\alpha_1 = a_1 + 0.382(b_1 - a_1) \text{ and } \beta_1 = a_1 + 0.618(b_1 - a_1)$$

The values of the function $f(\alpha_1)$ and $f(\beta_1)$ are calculated. The counter of the iterations k is set to one.

Step 1 Stop test. If $(b_k - a_k) < \varepsilon$ stop and the solution of the problem is

$$\theta^* = \frac{a_k + b_k}{2}$$

Otherwise, if $f(\alpha_k)$ is greater than $f(\beta_k)$ go to step 2 (Case 2 in Fig. A.4.2). If it is less, go to step 3 (Case 1 in Fig. A.4.2).

Step 2 Let $a_{k+1} = \alpha_k$ and $b_{k+1} = b_k$.

By definition of the golden section, we have

$$\alpha_{k+1}=\beta_k, f(\alpha_{k+1}) = f(\beta_k), \text{ and } \beta_{k+1}= a_{k+1} + 0.618(b_{k+1} - a_{k+1}).$$

Compute $f(\beta_{k+1})$ and go to step 4.

Step 3 Let $a_{k+1} = a_k$ and $b_{k+1} = \beta_k$.

Furthermore, it results

$$\beta_{k+1} = \alpha_k, f(\beta_{k+1}) = f(\alpha_k) \text{ and } \alpha_{k+1} = a_{k+1} + 0.382(b_{k+1} - a_{k+1}).$$

Compute $f(\alpha_{k+1})$ and go to step 4.

Step 4 Update the counter $k = k + 1$ and repeat from step 1.

As an example, Fig. A.4.3 reports the relevant variables of the golden section method for the following problem:

$$\min_{-3 \le \theta \le 5}(\theta^2 + 2\theta) \tag{A.4.7}$$

with stop threshold $\varepsilon = 0.2$.

k	a_k	b_k	α_k	β_k	$f(\alpha_k)$	$f(\beta_k)$	ε
1	-3.00	5.00	-0.104	2.104	-0.197184	8.634816	8.00
2	-3.00	2.10	-1.152	-0.104	-0.976789	-0.197184	5.10
3	-3.00	-0.10	-1.952	-1.152	-0.094366	-0.976789	2.90
4	-1.95	-0.10	-1.152	-0.773	-0.976789	-0.948402	1.85
5	-1.95	-0.77	-1.525	-1.152	-0.724456	-0.976789	1.18
6	-1.52	-0.77	-1.152	-1.045	-0.976789	-0.997966	0.75
7	-1.15	-0.77	-1.045	-0.910	-0.997966	-0.991941	0.38
8	-1.15	-0.91	-1.065	-1.045	-0.995813	-0.997966	0.24
9	-1.06	-0.91	-1.045	-0.966	-0.997966	-0.998854	0.15
10	-1.05	-0.91	-0.966	-0.959	-0.998854	-0.998323	0.13
11	-1.05	-0.96	-1.014	-0.966	-0.999805	-0.998854	0.09
12	-1.05	-0.97	-1.017	-1.014	-0.999727	-0.999805	0.08
13	-1.02	-0.97	-1.014	-0.984	-0.999805	-0.999756	0.05
14	-1.02	-0.98	-1.005	-1.014	-0.999976	-0.999805	0.03
15	-1.02	-1.01	-1.016	-1.005	-0.999757	-0.999976	0.00

Fig. A.4.3 Relevant variables of the golden section algorithm.

The algorithms discussed so far are based exclusively on evaluations of the objective function in different points without using derivatives. If the function is differentiable, and especially if its derivative $f'(\theta)$ is easily calculated, algorithms exploiting the "information" contained in the derivative can be used to solve the problem (A.4.7). The main steps of the *bisection algorithm* which halves the uncertainty interval at each iteration on the basis of the value assumed by the

derivative at the mid point of the current uncertainty interval are described in the following. The algorithm is asymptotically convergent if the function $f(\theta)$ is convex (also in a weaker sense than that introduced in section A.1.3).

Bisection algorithm

Step 0 Initialization. The maximum width, ε, of the final interval of uncertainty is chosen. If the function is convex, the condition $f'(\theta_{min}) \geq 0$ implies that θ_{min} is a minimum point, analogously $f'(\theta_{max}) \leq 0$ implies that θ_{max} is a minimum point. Otherwise set the extremes of the initial interval $a_1 = \theta_{min}$, $b_1 = \theta_{max}$ and the counter at $k = 1$.

Step 1 The derivative $f'(\theta_k)$ is calculated at the mid-point of the uncertainty interval $\theta_k = 1/2(a_k + b_k)$. If $f'(\theta_k) = 0$, the solution is $\theta^* = \theta_k$. If $f'(\theta_k) > 0$ go to step 2, if $f'(\theta_k) < 0$, to step 3.

Step 2 Let $a_{k+1} = a_k$, $b_{k+1} = \theta_k$ and go to step 4.

Step 3 Let $a_{k+1} = \theta_k$, $b_{k+1} = b_k$ and go to step 4.

Step 4 Stop test. If $b_{k+1} - a_{k+1} < \varepsilon$ stop and the solution of the problem is:

$$\theta^* = \frac{a_{k+1} + b_{k+1}}{2} = \theta_{k+1}$$

Otherwise increase the counter, $k=k+1$, and repeat from step 1.

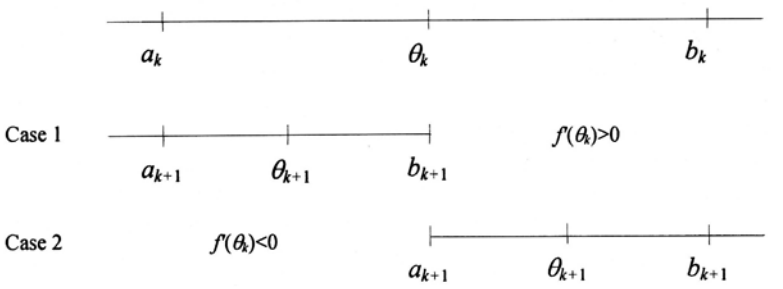
Fig. A.4.4 illustrates the bisection algorithm for the problem (A.4.7).

A.4.2.2. Unconstrained multi-dimensional optimization algorithms

The unconstrained multi-dimensional optimization problem:

$$\min_{x \in E^n} f(\boldsymbol{x}) \tag{A.4.8}$$

can be solved by using different algorithms, some of which are based exclusively on the calculation of the values of the objective function, others on the use of first and second order derivatives.



k	a_k	b_k	θ_k	$f(\theta_k)$	$f'(\theta_k)$	ε
1	-3.00	5.00	1.00	3.00	4.00	8.00
2	-3.00	1.00	-1.00	-1.00	0.00	4.00
3	-1.00	1.00	0.00	0.00	2.00	2.00
4	-1.00	0.00	-0.50	-0.75	1.00	1.00
5	-1.00	-0.50	-0.75	-0.94	0.50	0.50
6	-1.00	-0.75	-0.88	-0.98	0.25	0.25
7	-1.00	-0.88	-0.94	-1.00	0.13	0.13
8	-1.00	-0.94	-0.97	-1.00	0.06	0.06
9	-1.00	-0.97	-0.98	-1.00	0.03	0.03
10	-1.00	-0.98	-0.99	-1.00	0.02	0.02
11	-1.00	-0.99	-1.00	-1.00	0.01	0.01

Fig. A.4.4 Illustration of the bisection algorithm.

A brief description of some *descent direction algorithms* follows. These algorithms make use of the results described in section A.1, and, at each iteration k, search for the minimum of the function $f(\boldsymbol{x})$ along a direction of negative directional derivative $\boldsymbol{h}^k$ (linear minimization). The algorithms converge towards a null gradient point (stationariety point) of the function $f(\boldsymbol{x})$; they converge towards a global minimum point if the objective function is convex. The simplest of such algorithms, known as gradient algorithm, assumes the opposite of the gradient as descent direction. The main steps of the algorithm are given below.

Gradient algorithm

Step 0 Initialization. The stop parameter ε is fixed. This can be either the maximum gradient module or the maximum deviation between the values of $f(\boldsymbol{x})$ in two successive iterations. An initial point $\boldsymbol{x}^1$ is chosen and the counter of the iteration k is set to one.

Step 1 Calculation of the search direction

$$\boldsymbol{h}^k = -\nabla f(\boldsymbol{x}^k)$$

Step 2 Line search. The value of the parameter θ minimizing the function of a single variable $f(\boldsymbol{x}_k + \theta \boldsymbol{h}_k)$ is sought

$$\theta^k = \underset{0 \le \theta \le \theta^*}{argmin} f(\boldsymbol{x}^k + \theta \boldsymbol{h}^k)$$

where θ^* is a prefixed, large enough value. The line search can be carried out by using one of the algorithms described in the previous section.

Step 3 Calculation of the next point as

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \theta^k \boldsymbol{h}^k$$

Step 4 Stop test. If the module of the function gradient in $\boldsymbol{x}^{k+1}$ is less than the stop threshold:

$$||\nabla f(\boldsymbol{x}^{k+1})|| < \varepsilon$$

or if the relative difference of two successive values of the objective function is less than the stop threshold:

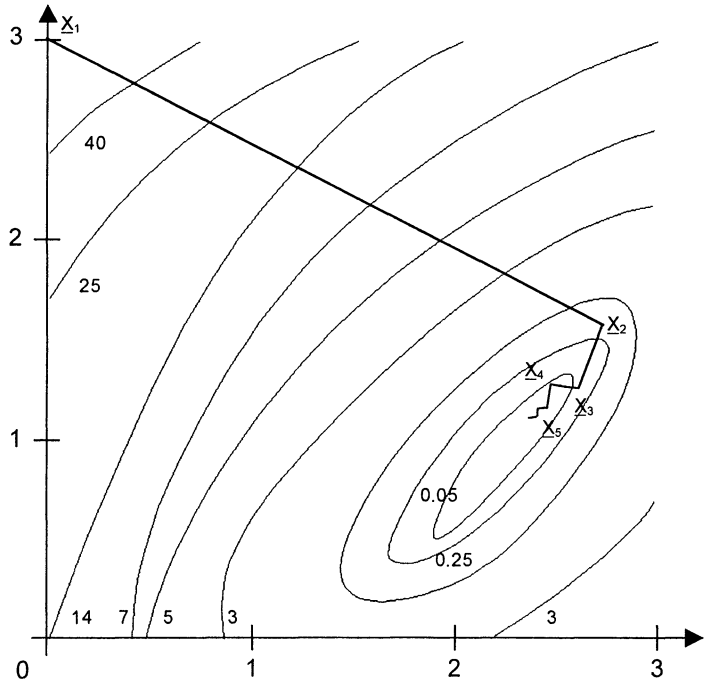
$$\frac{\left[f\left(\boldsymbol{x}^k\right) - f\left(\boldsymbol{x}^{k+1}\right)\right]}{f\left(\boldsymbol{x}^k\right)} < \varepsilon$$

stop, otherwise increase the counter ($k = k + 1$) and repeat from step 1.

Fig. A.4.5 describes an application of the gradient algorihm to the minimization of the function:

$$f = (x_1 - 2)^4 + (x_1 - 2\,x_2)^2 \qquad \text{(A.4.9)}$$

with stop parameter $\varepsilon = 0.10$.



k	x_k $f(x_k)$	$\nabla f(x_k)$	$\|\nabla f(x_k)\|$	$h_k=-\nabla f(x_k)$	θ^k	x_{k+1}
1	(0.00,3.00) 52.00	(-44.00,24.00)	50.12	(44.00,-24.00)	0.062	(2.70,1.51)
2	(2.70,1.51) 0.34	(0.73,1.28)	1.47	(-0.73,-1.28)	0.24	(2.52,1.20)
3	(2.52,1.20) 0.09	(0.80,-0.48)	0.93	(-0.80,0.48)	0.11	(2.43,1.25)
4	(2.43,1.25) 0.04	(0.18,0.28)	0.33	(-0.18,-0.28)	0.31	(2.37,1.16)
5	(2.37,1.16) 0.02	(0.30,-0.20)	0.36	(-0.30,0.20)	0.12	(2.33,1.18)
6	(2.33,1.18) 0.01	(0.08,0.12)	0.14	(-0.08,-0.12)	0.36	(2.30,1.14)
7	(2.30,1.14) 0.009	(0.15,-0.08)	0.17	(-0.15,0.08)	0.13	(2.28,1.15)
8	(2.28,1.15) 0.007	(0.05,0.08)	0.09			

Fig. A.4.5 Graphic representation and relevant variables for an application of the gradient algorithm.

This figure shows a typical characteristic of the gradient algorithm: in the first iterations a rapid decrease in the objective function is observed, while successive iterations show smaller reductions and zigzagging towards the optimum value. The problem (A.4.8) can be solved with other algorithms whose structure is substantially similar to that described above, apart from the calculation of the descent direction $\boldsymbol{h}_k$. These algorithms, in order to accelerate convergence, use directions obtained by "deflecting" the gradient and for this reason they are denoted "*deflected gradient*" *algorithms*.

Fletcher and Reevers' conjugate gradient algorithm deflects the opposite gradient at each iteration, adding a positive multiple of the direction used in the previous iteration. In the case of quadratic objective function ($f(\boldsymbol{x})=\boldsymbol{x}^T \boldsymbol{H}\, \boldsymbol{x}$), this algorithm generates a series of conjugate directions (from which it derives its name) with respect to the matrix $\boldsymbol{H}$, and converges at the optimum point in a finite number of iterations equal to the number of components of $\boldsymbol{x}$. In the general case, it usually converges more quickly than the gradient algorithm, and in particular solves the zigzagging problems in proximity of the minimum point typical of the gradient.

The description of the conjugate gradient algorithm is basically similar to that of the gradient algorithm. The only difference is in the calculation of the descent direction (Step 1) which is substituted as follows.

Conjugate gradient algorithm

Step 1 Calculation of the search direction

$$\boldsymbol{h}^k = -\nabla f(\boldsymbol{x}^k) + \alpha_k \boldsymbol{h}^{k-1}$$

$$\alpha_k = \frac{\left\|\nabla f\left(\boldsymbol{x}^k\right)\right\|^2}{\left\|\nabla f\left(\boldsymbol{x}^{k-1}\right)\right\|^2}$$

A.4.2.3. Bounded variables multi-dimensional optimization algorithms

The problem of minimizing the objective function, imposing constraints on the lower and/or upper bounds of the components of the vector $\boldsymbol{x}$ is slightly more complex than that of unconstrained optimization (A.4.8). In this case the constraint $g_i(\boldsymbol{x})$ can be written as:

$$x_i \geq c_i \text{ and/or } x_i \leq c_i$$

The variables x_i can be easily modified so that the constraints are always expressed in the form $x_i \geq 0$, therefore the problem of optimization with inequality constraints can be formally expressed as:

$$\min_{x \geq 0} f(\boldsymbol{x}) \tag{A.4.10}$$

The problem (A.4.10) can be solved by using a feasible directions algorithm similar to those described previously. The main difference is that the descent direction used for the unconstrained problem (A.4.8), e.g. the opposite gradient, is not necessarily a feasible direction with respect to the feasibility set defined by the constraints of the problem. To solve this inconvenience when it occurs, the descent direction can be "projected" over the feasibility set as in the projected gradient algorithm described below.

Projected gradient algorithm

Step 0 Initialization. The stop parameter ε is fixed, a feasible initial point $\boldsymbol{x}_1$ is chosen (e.g. $\boldsymbol{x}_1 = \boldsymbol{0}$), the value of the objective function $f(\boldsymbol{x}_1)$ is calculated and the iterations counter k set to one.

Step 1 Calculation of the search direction. The components of the direction $\boldsymbol{h}^k$ are equal to the components of the gradient with changed sign if these components are feasible (i.e., if the x^k_i is positive and/or the gradient component is negative). Vice versa if the j-th component of the gradient changed of sign is not feasible, the corresponding component of h^k_i is set to zero; this corresponds to the projection of $-\nabla f(\boldsymbol{x}^k)$ over the hyperplane perpendicular to the j-th axis.

$$h_i^k = -\frac{\partial f(x^k)}{\partial x_i} \qquad \text{if } x_i^k > 0 \text{ and/or } -\frac{\partial f(x^k)}{\partial x_i} \geq 0$$

$$h_i^k = 0 \qquad \text{otherwise}$$

Step 2 Mono-dimensional search. The minimum of the function $f(\boldsymbol{x}^k + \theta\, \boldsymbol{h}^k)$ is searched for in the interval $[0, \theta^*]$ where θ^* is the maximum value allowing non-exit from the feasibility set (i.e. ensuring the non-negativity of all the components of x^{k+1}):

$$\theta^k = \underset{0 \leq \theta \leq \theta^*}{argmin}\, f(\boldsymbol{x}^k + \theta\, \boldsymbol{h}^k)$$

with $\theta^* = max_i \dfrac{x_i^k}{-h_i^k}$ for $i : h^k_i < 0$, otherwise $\theta^* = \infty$

(Note that for $h_i^k < 0$ it must result $x^k_i > 0$ because of step 1)

Step 3 Calculation of the next point

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \theta^k \boldsymbol{h}^k$$

Step 4 Stop test. This can be carried out on the projected gradient module

$$\left\| \boldsymbol{h}^k \right\| < \varepsilon$$

verifying the impossibility any further move along the projected gradient or heuristically, on the percentage decrease of the objective function in the last two iterations:

$$\left| \frac{f\left(\boldsymbol{x}^{k+1}\right) - f\left(\boldsymbol{x}^k\right)}{f\left(\boldsymbol{x}^k\right)} \right| < \varepsilon$$

Otherwise increase the counter, $k = k+1$, and repeat from step 1.

A.4.2.4. Linearly constrained multi-dimensional optimization algorithms

The problem of minimizing the objective function over a closed set defined by linear inequality and/or equality constraints can be stated formally as:

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\boldsymbol{A}\,\boldsymbol{x} \le \boldsymbol{a} \\ &\boldsymbol{B}\,\boldsymbol{x} = \boldsymbol{b} \end{aligned} \qquad \text{(A.4.11)}$$

This problem can be solved with different algorithms which differ in the way they generate the "feasible descent direction" $\boldsymbol{h}_k$, i.e. a direction along which it is possible to move while reducing the objective function $f(\boldsymbol{x})$ and remaining within the set S defined by the constraints. A description of the Frank-Wolfe algorithm follows which at each iteration generates the direction $\boldsymbol{h}_k$, minimizing a linear approximation of $f(\boldsymbol{x})$.

Frank-Wolfe algorithm

Step 0 Initialization. The stop parameter ε is fixed; a feasible initial point $\boldsymbol{x}^1$ is chosen and the iteration counter k is set to one.

Step 1 Generation of the feasible direction. The linear programming problem is solved:

$$\begin{aligned} &\boldsymbol{y}^k = argmin\ \nabla f(\boldsymbol{x}^k)^T \boldsymbol{y} \\ &\boldsymbol{A}\,\boldsymbol{y} = \boldsymbol{a} \\ &\boldsymbol{B}\,\boldsymbol{y} \le \boldsymbol{b} \end{aligned} \qquad \text{(A.4.12)}$$

The problem is equivalent to the minimization of the linear approximation of $f(\boldsymbol{x})$ at the point $\boldsymbol{x}^k$ given by:

$$f_L(\boldsymbol{y}) = f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^T (\boldsymbol{y} - \boldsymbol{x}^k)$$

once the constant terms are eliminated. Problem (A.4.12) can be solved with the simplex algorithm or one of its variants[6]. The descent direction is $\boldsymbol{h}^k = \boldsymbol{y}^k - \boldsymbol{x}^k$

Step 2 Line search. The linear minimum of the function $f(\boldsymbol{x}^k + \theta\,\boldsymbol{h}^k)$ is searched for:

$$\theta^k = \underset{0 \le \theta \le 1}{argmin}\, f(\boldsymbol{x}^k + \theta\,\boldsymbol{h}^k)$$

for θ included in the interval [0,1]. The points $\boldsymbol{x}^k$ and $\boldsymbol{y}^k$ correspond to the extreme values of θ; since both are feasible by construction and the set S is convex, all the points of the segment that joining them are feasible.

Step 3 Calculation of the next point:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \theta^k \boldsymbol{h}^k$$

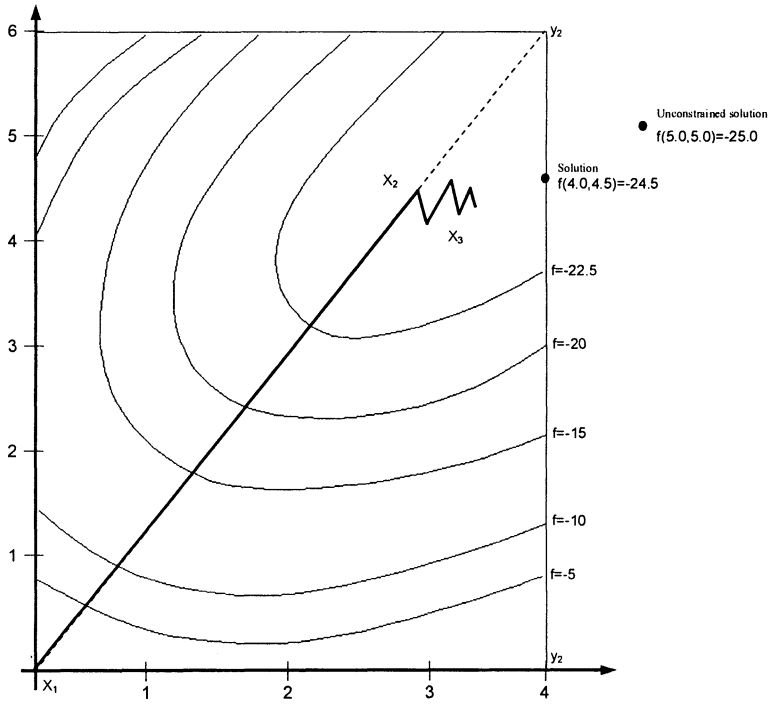
Step 4 Stop test. If $\nabla f(\boldsymbol{x}^k)^T(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) > -\varepsilon$ or, more simply (but less effectively), $|f(\boldsymbol{x}^k) - f(\boldsymbol{x}^{k+1})| \,/\, f(\boldsymbol{x}^k) < \varepsilon$ stop; otherwise increase the counter, $k = k + 1$, and repeat from step 1.

Also in this case it is possible to demonstrate that if $f(\boldsymbol{x})$ is a convex function, the algorithm converges to the solution of the problem (A.4.11).

Fig. A.4.6 illustrates the application of the Frank-Wolfe algorithm to the following optimum problem:

$$\begin{aligned} &min\ x^2_1 + 2\,x^2_2 - 2\,x_1\,x_2 - 10\,x_2 \\ &0 \le x_1 \le 4 \\ &0 \le x_2 \le 6 \end{aligned}$$

with stop parameter $\varepsilon = 0.10$.



k	$\nabla f(x^k)$	y^k	θ^k	x^{k+1}	$f(x^{k+1})$	$f(x^k)$-$f(x^{k+1})$
0				0.000		
				0.000		
1	0.000	(4,6)	0.750	3.000	-22.500	22.500
	-10.000			4.500		
2	-3.000	(4,0)	0.119	3.119	-23.213	0.713
	2.000			3.969		
3	-1.700	(4,6)	0.206	3.301	-23.446	0.233
	-0.362			4.385		
4	-2.168	(4,0)	0.063	3.344	-23.622	0.176
	0.938			4.111		
5	-1.534	(4,6)	0.144	3.439	-23.728	0.106
	-0.244			4.383		
6	-1.888	(4,0)	0.045	3.464	-23.816	0.089
	0.654			4.186		

Fig. A.4.6 Graphic representation and significant variables for an application of the Frank-Wolfe algorithm.

A.5. Variational inequality problems

Let $\varphi(x)$ be a vectorial function of a vector defined in a convex set $S \subseteq E^n$, with values in the set $T = \varphi(S) = \{\varphi(x) : x \in S\} \subseteq E^n$; the mathematical problem, called variational inequality, with solution in the point $x^* \in S$ is defined as:

$$\varphi(x^*)^T (x - x^*) \geq 0 \quad \forall x \in S \tag{A.5.1}$$

In other words, the problem of the variational inequality of a vectorial function of a vector consists of the search for point x^* at which the vector function $\varphi(x^*)$ has a non-negative scalar product (i.e. angles $\leq \pi/2$) with all the vectors joining the point x^* with every other point x of the set of definition S.

Variational inequality problems can be considered a generalization of minimization problems, in particular of the conditions of virtual minimum (A.4.5b), since the vectorial function of vector $\varphi(x)$ is not required to be the gradient of a scalar function of vector $f(x)$. To show this, let consider the generic minimization problem:

$$\begin{aligned} & x^* = argmin\, f(x) \\ & x \in S \end{aligned} \tag{A.5.2}$$

If the function $f(x)$ is differentiable, its gradient $\nabla f(x)$ is a vectorial function of vector, and the virtual minimum conditions are given by:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in S \tag{A.5.3}$$

It then results that the variational inequality (A.5.1) in the function $\varphi(x) = \nabla f(x)$ coincides with the expression of the virtual minimum conditions of the minimization problem (A.5.2). Furthermore, if the gradient $\nabla f(x)$ exists, the minimization problem (A.5.2) can be reformulated as:

$$\begin{aligned} & x^* = argmin\, z(x) = \int_0^x \nabla f(t)^T\, dt \\ & x \in S \end{aligned} \tag{A.5.4}$$

On the other hand, given a vectorial function of vector $\varphi(x)$ with symmetrical Jacobian $\boldsymbol{Jac}[\varphi(x)]$, a minimization problem can be defined:

$$\begin{aligned} & x^* = argmin\, f(x) = \int_0^x \varphi(t)^T\, dt \\ & x \in S \end{aligned} \tag{A.5.5}$$

In general, the value of the curvilinear integral appearing in (A.5.4) depends on the integration path; however, if the Jacobian $\boldsymbol{Jac}[\varphi(x)]$ of the integrating function $\varphi(x)$ is symmetric, the value of the integral is independent of the integration path,

being the set of definition convex (Green's theorem). In other words, if and only if the integrating function $\varphi(x)$ has symmetrical Jacobian, the former can be the gradient of a function $f(x)$, i.e. $\nabla f(x) = \varphi(x)$, of which $\boldsymbol{Jac}[\varphi(x)]$ is the (symmetrical) Hessian matrix. In the equivalent minimization problem (A.5.5) is correctly defined, the necessary conditions of virtual minimum are given by:

$$\nabla f(x^*)^T (x - x^*) \geq 0 \quad \forall x \in S$$

i.e. (A.5.6)

$$\varphi(x^*)^T (x - x^*) \geq 0 \quad \forall x \in S$$

It can immediately be seen that the condition (A.5.6) is formally coincident with the variational inequality (A.5.1).

If the function $\varphi(x)$ is continuous and differentiable with symmetrical and semi-definite positive Jacobian $\boldsymbol{Jac}[\varphi(x)]$, a vector x^* solving the constrained optimization model (A.5.5) solves the corresponding variational inequality (A.5.1) and vice versa.

In this case, in fact, the objective function of the problem (A.5.5) $f(x)$ is differentiable with continuous gradient and continuous positive semi-definite Hessian matrix, since $\nabla f(x) = \varphi(x)$, and $\boldsymbol{Hess}[f(x)] = \boldsymbol{Jac}[\varphi(x)]$, therefore $f(x)$ is convex, and so the conditions of virtual minimum (A.5.6) are necessary and sufficient.

A.5.1. Properties of variational inequalities

Sufficient conditions for the existence of at least one solution of the variational inequality (A.5.1) can be obtained by applying Brouwer's theorem, as follows.

Theorem. The variational inequality problem (A.5.1) has at least one solution if:

S is a nonempty, compact and convex set;
$\varphi(x)$ is a continuous function.

Sufficient conditions for the uniqueness of the variational inequality solution are given by the following theorem.

Theorem. The variational inequality (A.5.1) has at most one solution if:

$\varphi(x)$ is a strictly monotone increasing function, i.e.:

$$(\varphi(x') - \varphi(x''))^T (x' - x'') > 0 \qquad \forall x', x'' \in S$$

In fact, if there existed two different vectors solving the variational inequality, $x_1^* \neq x_2^* \in S$, we would have:

$$\varphi(x_1^*)^T(x - x_1^*) \geq 0 \quad \forall x \in S_x \tag{A.5.7a}$$
$$\varphi(x_2^*)^T(x - x_2^*) \geq 0 \quad \forall x \in S_x \tag{A.5.7b}$$

From (A.5.7a) for $x = x_2^*$:

$$\varphi(x_1^*)^T(x_2^* - x_1^*) \geq 0 \tag{A.5.8a}$$

Furthermore, from (A.5.7a) for $x = x_1^*$, it results:

$$\varphi(x_2^*)^T(x_1^* - x_2^*) \geq 0$$

i.e. (A.5.8b)

$$-\varphi(x_2^*)^T(x_2^* - x_1^*) \geq 0$$

Adding (A.5.8a) and (A.5.8b), it would follow:

$$(\varphi(x_1^*) - \varphi(x_2^*))^T(x_2^* - x_1^*) \geq 0$$

i.e.

$$(\varphi(x_2^*) - \varphi(x_1^*))^T(x_2^* - x_1^*) \leq 0$$

which contradicts the monotonicity assumption.

A.5.2. Solution algorithms for variational inequality problems

Solution algorithms for variational inequality problem (A.5.1), in the case of the function $\varphi(x)$ with symmetrical Jacobian, are based on algorithms solving the equivalent minimization problem (A.5.5) described in section A.4.2. Note that in this case the gradient, $\nabla f(x)$, of the objective function of the minimization problem, used by the algorithm, is given by the function $\varphi(x)$ defining the variational inequality.

In the general case of a function $\varphi(x)$ with non-symmetrical Jacobian, various solution algorithms can be adopted; even though their convergence analysis usually requires conditions that are not easily verifiable. One of the simplest, called diagonalization algorithm, generates a succession of vectors, x^k , starting from a feasible point, $x^0 \in S$, solving a succession of variational inequalities defined by functions with diagonal Jacobians approximating the problem (A.5.1). In particular, at a point $x^* \in S$, the i-th component function $\varphi_i(x)$ of the vectorial function $\varphi(x)$, can be approximated by a function $\varphi^*_i(x_i)$ obtained by fixing all the other component of x to their values x_j^* (i.e. diagonalizing the Jacobian):

$$\varphi_i(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots) \cong \varphi^*_i(x^*_1, \ldots, x^*_{i-1}, x_i, x^*_{i+1}, \ldots) = \varphi^*_i(x_i) \; \forall i$$

Thus the variational inequalilty (A.5.1) can be approximated by a variational inequality defined by a function of $\varphi^*(x)$ with diagonal Jacobian:

$$\varphi(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \cong \Sigma_i \varphi^*_i(x_i) (x_i - x_i^*) \geq 0 \quad \forall \boldsymbol{x} \in S_x \tag{A.5.9}$$

The solution of the approximate variational inequality (A.5.9) can be obtained by solving the equivalent minimization problem (A.5.5), with one of the algorithms described in section A.4.2.

Notes

(1) Two mathematical problem are said to be equivalent if the solutions of one problem are also solutions of the other and vice versa. In this case the analysis of the theoretical properties of the solutions such as their existence and uniqueness, and the convergence analysis of resolutive algorithms can be carried out for only one of the two problems.

(2) Banach's theorem requires the function $\psi(x)$, defined over S with values in $T \subseteq S$, to be a contraction (a stricter property than that of monotonicity) over a complete set (a weaker property than that of compactness), or that the function $\psi(x)$ is a quasi-contraction (implying monotonicity) over a compact set. Note that in both cases the function is continuous.

(3) More in general it is sufficient to have:

$$(\psi(x') - \psi(x''))^T (x' - x'') < (x' - x'')^T (x' - x'') \quad \forall x' \neq x'' \in S$$

(4) If the function $\psi(x)$ is the realization of a random variable, and an unbiased estimate of its value is available, the convergence is almost certain. (This is the most general of the cases originally analysed in Blum's theorem).

(5) The results reported and the algorithms described can easily be extended to maximum points, bearing in mind that the maximum points of a function correspond to the minimum points of the opposite function $-f(x)$.

(6) The solution of the problem A.4.12 is generally one of the vertices of the set defined by the linear equations and inequalities. Therefore the Frank-Wolf algorithm can move only along directions pointing to the vertices and presents zigzagging problems in proximity of the minimum similar to those described for the gradient algorithm.

References

[1] Abdulaal M., and L.J. Le Blanc (1979). *Continuous equilibrium network design models.* Transportation Research 13 B: 19-32.

[2] Abkowitz M.D. (1981). An analysis of the commuter departure time decision. Transportation 10: 283-297.

[3] Ahuja R.K. T.L. Magnanti, and J.B. Orlin (1993). *Networks flows: Theory, Algorithms, and Application.* Prentice –Hall, Englewood Cliffs, NJ USA.

[4] Akcelik R. (1988). *The Highway Capacity Manual delay formula for signalized intersections.* ITE Journal: 23-27.

[5] Alexander E. R. (1997). *Introduzione alla Pianificazione: teorie, concetti e problemi attuali.* Translation of *Approaches to Planning: introduction to current planning theories* (by F. D. Moccia), Clean, Naples, Italy.

[6] Algers S., A. Daly, and S. Widlert (1993). *The Stockholm Model System.* Technical report, Hague Consulting Group, The Hague, The Netherlands.

[7] Antoniou C. (1997). *Demand simulation for dynamic traffic assignment.* SM Thesis, Department of Civil Engineering, Massachusetts Institute of Tecnology.

[8] Antonisse R.W., A. Daly, and H. Gunn (1986). *The primary destination tour approach to modelling trip chains.* Proceedings of Seminar M on Transportation Planning Methods at the 14th PTRC Summer Annual Meeting, University of Sussex, England.

[9] Ashok K. and Ben-Akiva M. (1993) *Dynamic Origin-destination Matrix Estimation and Prediction for Real Time-Traffic Management Systems.* in C. F. Daganzo editor, International Symposium on Transportation and Traffic Theory, Elsevier Science Pub. Co. : 465-484.

[10] Astarita V. (1996). *A continuous time link based model for dynamic network loading based on travel time function.* Proceedings of the 13th International Symposium on Transportation and Traffic Theory. Lyon, France.

[11] Bath C. (1997). *Recent Methodological Advances Relevant to Activity and Travel Behavior Analysis.* Proceedings of the VIII IATBR conference, Resource papers, Austin, Texas.

[12] Bayliss B. (1988). *The Measurement of Supply and Demand in Freight Transport.* Hants, England, Avebury Grower Publishing Company Ltd.

[13] Beckman M., C.B. McGuire, and Winsten C.B. (1956). *Studies in the economics of transportation* Yale University Press, New Haven, CT.

[14] Bell M. (1995). *Stochastic user equilibrium assignment in networks with queues.* Transportation Research 29B : 125-137.

[15] Bell M., D. Inaudi, W. Lam, and G. Ploss (1993). *Stochastic User Equilibrium Assignment and Iterative Balancing* Proceedings of the 12th International Symposium on Traffic and Transportation Theory, Berkekey: 427-440.

[16] Ben Akiva M., and B. Boccara (1995). *Discrete choice models with latent choice sets.* International Journal of Research in Marketing 12: 9-24.

[17] Ben Akiva M., and B. Francois (1983). *μ Homogeneous Generalized Extreme Value Model.* Working paper, Department of Civil Engineering, MIT Cambridge, Mass.

[18] Ben Akiva M., and D. Bolduc (1996). *Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariance Structure.* Working Paper, Department of Economics, MIT, Boston, Mass.

[19] Ben Akiva M., and S. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, Cambridge, Mass.

[20] Ben Akiva M., and T. Atherton (1977). *Methodology for Short-Range Travel Demand Predictions.* Journal of Transport Economics and Policy 11: 224-261.

[21] Ben Akiva M., J. Bowman, and D. Gopinath (1996). *Travel Demand Model System for the Information Era.* Transportation 23: 241-266.

[22] Ben Akiva M., M. Cyna, and A. de Palma (1984). *Dynamic model of peak period congestion.* Transportation Research 18B: 339-355.

[23] Ben Akiva M., M.J. Bergman, A.L. Day, and R. Ramaswamy (1984). *Modelling interurban route choice behaviour.* Proceedings of the 9th International Symposium on Transportation and Traffic Theory, VNU Science Press: 299-330.

[24] Ben-Akiva M., and Morikawa T. (1990). *Estimation of switching models from Revealed Preference and Stated Intention.* Transportation Research 24 A.

[25] Ben-Akiva M., M. Bierlaire, J. Bottom, H. Koutsopoulos, and R. Mishalani (1997). *Development of a route guidance generation system for real-time application.* Proceedings of the 8th IFAC Symposium on Transportation Systems, Chania, Greece.

[26] Ben-Akiva, M., and M. Bierlaire (1999). *Discrete choice methods and their application to short term travel decisions.* Handbook of Transportation Science, R.W. Hall ed., Kluwer Academic Publishers: 5-33.

[27] Bergman L. (1990). *The development of computable general equilibrium modeling.* in General equilibrium modelling and economic policy analysis (Bergman L., Jorgenson Dale W., Zalai E. ed.) Blackwell, Oxford.

[28] Bernstein D. and Smith T. E. (1994). Equilibria for Networks with Lower Semicontinuous Costs: With an Application to Congestion Pricing. *Transportation Science* 28: 221–235.

[29] Bianco L. (1986). *The role of quantitative methods in urban transportation planning.* Proceedings of the International Seminar "The management and planning of urban transport systems: from theory to practice, Montreal, Canada.

[30] Bifulco G.N. (1993). *A stochastic user equilibrium assignment model for the evaluation of parking policies.* EJOR 71: 269-287.

[31] Biggiero L. (1991). *Un modello Comportamentale per la Generazione degli Spostamenti non sistematici in area urbana.* Trasporti e Trazione 4.

[32] Biggiero L., and Postorino M.N., (1995), *La calibrazione di modelli di scelta modale mediante l'uso congiunto di dati RP e SP.* In Cascetta E., and Salerno G., *Sviluppi della ricerca sui sistemi di Trasporto.* Ed. Franco Angeli, Rome.

[33] Billheimer J. W., and P. Gray (1973). *Network design with fixed and variable cost elements.* Transportation Science 7: 49-74.

[34] Blum J.R. (1954). Multidimensional Stochastic Approximation Methods. Ann. Math. Stat. 25: 737–744.

[35] Bolduc, D., B. Fortin, and M.A. Fournier (1996). "The Impact of Incentive Policies to Influence Practice Location of General Practitioners: A Multinomial Probit Analysis", Journal of Labor Economics 14: 703-732.

[36] Bouzaiene-Ayari B., Gendreau M., and Nguyen S. (1997). "Transit Equilibrium assignment Problem: A Fixed-Point Simplicial -Decomposition Solution Algorithm", Operations Research.

[37] Bouzaiene-Ayari B., M. Gendrau, and S. Nguyen (1998). *Passenger assignment in congested transit networks: a historical perspective.* In P. Marcotte and S. Nguyen (eds.) "Equilibrium and advanced transportation modelling", Kluwer: 47-71.

[38] Bouzaiene-Ayari B., M. Gendreau , and S. Nguyen (1995). "On the Modelling of Bus Stops in Transit Networks", Centre de recherche sur les transports, Université de Montréal.

[39] Box G., Hunter W., and Hunter G. (1978), Statistics for experiment, John Wiley & S.

[40] Boyce D. E., and B. N. Janson (1980). *A discrete transportation network design problem with combined trip distribution and assignment.* Transportation Research 14 B: 147-154.

[41] Boyce D.E., B. Ran, and L. J. Le Blanc (1991). *Dynamic user-optimal traffic assignment model: a new model and solution technique.* Proceedings of the 1st TRISTAN, Montreal, Canada.

[42] Boyer K. D. (1998). *Principles of transportation economics.* Addison Wesley, Longman.

[43] Bradley M.A., and Daly A.J. (1992). *Estimation of logit choice models using mixed Stated Preference and Revealed Preference information.* 20th PTRC Sam, Manchester.

[44] Branston D. (1976). *Link Capacity Functions: a review.* Transportation Research 10: 223-236.

[45] Brog W., Ampt E., (1982). *State of the art in the collection of travel behaviour data.* in *Travel Behaviour for the 1980's*, Special Report 201, National Research Council, Washington, DC.

[46] Burrell J.E. (1968). *Multiple route assignment and its application to capacity restraint.* In Proceedings o the 4t International Symposium on the Theory of Road Traffic Flow, W. Leutzbach and P. Baron eds. Karlsruhe, Germany.

[47] Cantarella G. E., E. Cascetta, V. Adamo, and V. Astarita (1999). *A doubly dynamic traffic assignment model for planning applications.* Proceedings of the 14th International Symposium on Transportation and Traffic Theory, Jerusalem, Israel.

[48] Cantarella G. E., G. Improta, and A. Sforza (1991). *Road network signal setting: equilibrium conditions.* In M. Papageorgiou ed. "Concise encyclopedia of traffic and transportation systems", Pergamon Press: 366-371.

[49] Cantarella G.E. (1997). *A General Fixed-Point Approach to Multi-Mode Multi-User Equilibrium Assignment with Elastic Demand.* Transportation Science 31: 107-128.

[50] Cantarella G.E., and A. Vitetta (2000). *Stochastic assignment to high frequency transit networks: models, and algorithms, and applications with different perceived cost distributions.* In Proceedings of the 7th Meeting of the EURO Working Group on Transportation. Helsinki, and Finland, and August 1999, forthcoming.

[51] Cantarella G.E., and E. Cascetta (1995). *Dynamic Processes and Equilibrium in Transportation Networks: Towards a Unifying Theory.* Transportation Science 29: 305–329.

[52] Cantarella G.E., and E. Cascetta (1998). *Stochastic Assignment to Transportation Networks: Models and Algorithms.* In Equilibrium and Advanced Transportation Modelling, P. Marcotte, S.Nguyen (editors): 87-107. (Proceedings of the International Colloquium, Montreal, Canada October, 1996).

[53] Cantarella G.E., and M. Binetti (1998). *Stochastic Equilibrium Traffic Assignment with Value-of-Time Distributed Among User.* In International Transactions of Operational Research 5: 541-553.

[54] Cantarella G.E., and M. Binetti (2000). *Stochastic Assignment with Gamma Distributed Perceived Costs.* Proceedings of the 6th Meeting of the EURO Working Group on Transportation. Gothenburg, and Sweden, and September 1998, forthcoming.

[55] Cantarella G.E., F.A. Viola, and A. Vitetta (1994). *Urban Network Design Through Multicriteria Analysis and Genetic Algorithms.* Proceedings of Seminar H on "Transportation Planning Methods" at the 22nd PTRC Summer Annual Meeting, University of Warwick, England.

[56] Cascetta E. (1984). *Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator.* Transportation Research 8B: 289-299.

[57] Cascetta E. (1986). *A class of travel demand estimators using traffic flows.* CRT Publication 375, Université de Montreal, Montreal, Canada.

[58] Cascetta E. (1987) *Static and dynamic models of stochastic assignment to transportation networks* in Flow control of congested networks (Szaego G., Bianco L., Odoni A. ed.), Springer Verlag, Berlin.

[59] Cascetta E. (1989) *A stochastic process approach to the analysis of temporal dynamics in transportation networks* Transportation Research 23B: 1-17.

[60] Cascetta E. (1995). *Modelli di scelta del percorso su reti di trasporto.* Proceedings of the seminary "Assegnazioni alle reti di trasporto" Department of Civil Engineering, University of Rome "Tor Vergata", Italy.

[61] Cascetta E. (1998). *Teoria e metodi dell'ingegneria dei sistemi di trasporto.* UTET, Turin, Italy.

[62] Cascetta E., A. Nuzzolo, and L. Biggiero (1992). *Analysis and modeling of commuters departure time and route choices in urban networks.* Proceedings of the 2nd International Seminar on Urban Traffic Networks, Capri, Italy.

[63] Cascetta E., A. Nuzzolo, and L. Biggiero (1995). *A system of behavioural models for the simulation of intercity travel demand in Italy.* Proceedings of the 7th WCTR, Sydney, Australia.

[64] Cascetta E., A. Nuzzolo, and V. Velardi (1994). *A time of the day tour based trip chaining model system for urban transportation planning.* in Transportation Planning Methods, Vol. I. Proceedings of the 22nd PTRC Summer Annual Meeting, University of Warwick, England.

[65] Cascetta E., A. Nuzzolo, F. Russo, and A. Vitetta (1996). *A new route choice logit model overcoming IIA problems: specification and some calibration results for interurban networks.* Proceedings of the 13th International Symposium on Transportation and Traffic Theory (Jean-Baptiste Lesort ed.), Pergamon Press.

[66] Cascetta E., and A. Nuzzolo (1982). *Analisi statistica del processo delle velocità in autostrada.* Autostrade 6.

[67] Cascetta E., and A. Nuzzolo (1986). *Uno schema comportamentale per la modellizzazione delle scelte di percorso nelle reti di trasporto pubblico urbano.* Proceedings of the IV PFT-CNR conference, Turin, Italy.

[68] Cascetta E., and A. Papola (2000). *A joint mode-run choice model to simulate the schedule influence at a regional level.* Proceedings of the 9th IATBR conference, Sidney, Australia.

[69] Cascetta E., and A. Papola (2000). *Implicit availability/perception models for the simulation of travel demand.* Transportation Research C, forthcoming.

[70] Cascetta E., and A. Papola. (2000). *Dominance among alternatives in choice set modeling: general theory and application to the destination choice.* Proceedings of the IX IATBR conference, Gold Coast, Queensland, Australia.

[71] Cascetta E., and B. Montella (1979). *Modelli di arrivo dei passeggeri alle fermate di un sistema di trasporto collettivo urbano.* La Rivista della Strada 453: 303-308.

[72] Cascetta E., and G.E. Cantarella (1991). *A day-to-day and within-day dynamic stochastic assignment model.* Transportation Research 25A: 277-291.

[73] Cascetta E., and Improta A.A. (1999). *Estimation of travel demand using traffic counts and other data sources.* Optimization Days 1999 (Michael Florian's special session) Montreal, Canada.

[74] Cascetta E., and P. Rostirolla (1989). *Un modello matematico per il calcolo della tariffa tecnico-economica.* Ingegneria Ferroviaria 6.

[75] Cascetta E., L. Biggiero, A. Nuzzolo, and F. Russo (1996). *A system of within-day dynamic demand and assignment models for scheduled inter-city*

services. Proceedings of the 24th European Transportation Forum, London, Great Britain.

[76] Cascetta E., M. Di Gangi, and G. Conigliaro (1996). *A Multi-Regional input-output model with elastic trade coefficients for the simulation of freight transport demand in Italy.* In Transportation Planning Methods, Proceedings of the 24th PTRC Summer Annual Meeting, England.

[77] Cascetta E., M. Gallo, and B. Montella (1998). *Optimal signal setting on traffic networks with stochastic equilibrium assignment.* Preprints of TRISTAN III - Puerto Rico - 17-23 June 1998.

[78] Cascetta E., M. Gallo, and B. Montella (1999). *An asymmetric SUE model for the combined assignment-control problem.* Selected proceedings of 8th WCTR, Volume 2, Pergamon: 189-202.

[79] Cascetta E., Nguyen S. (1986). *A unified framework for estimating or updating Origin-Destination matrices from traffic counts.* Transportation Research 22B: 437-455.

[80] Cascetta E., Postorino M.N. (2000). *Fixed point models for the estimation of O-D matrices using traffic counts on congested networks* submitted to Transportation Science.

[81] Cascetta E., Russo F. (1997). *Calibrating aggregate travel demand models with traffic counts: estimators and statistical performances.* Transportation 24: 271-293.

[82] Cascetta., Inaudi D. Marquis G. (1993). *Dynamic estimators of Origin-Destination matrices using traffic counts.* Transportation Science 27.

[83] Catling I. (1977). *A time-dependent approach to junction delays.* Traffic Engineering and Control 18.

[84] Chabini I., and S. Kachani (1999). *Analytical dynamic network loading models: analysis of a single link network.* Forthcoming in Transportation Research.

[85] Chankong V. and Y. Y. Haimes (1983). *Multiobjective decision making: theory and methodology.* Elsevier-North-Holland, New York.

[86] Chen M., and A. Sule Alfa (1991) *Algorithms for solving Fisk's Stochastic Traffic Assignment Model* Transportation Research 25B: 405-412.

[87] Chen M., and A.S. Alfa (1991). *A network design algorithm using a stochastic incremental traffic assignment approach.* Transportation Science 25: 215-224.

[88] Chen Y., and Florian M., (1995). *A coordinate Descent Method for the Bi-level O-D Matrix Adjustment Problem.* International Transactions in Operational Research 2: 165-179.

[89] Cochran W.G. (1963). *Sampling techniques.* John Wiley, New York.

[90] Costa P., and R. Roson (1988). *Transport margins, transportation industry and the multiregional economy. Some experiments with a model for Italy.* Ricerche Economiche 2: 237-287.

[91] Cremer M. and Keller H. (1987). *A new class of dynamic methods for the identification of Origin-Destination flows.* Transportation Research 21B: 117-132.

[92] Dafermos S.C. (1971). *An extended traffic assignment model with applications to two-way traffic* Transportation Science 5: 366-389.

[93] Dafermos S.C. (1972). *The Traffic Assignment Problem for Multi–Class User Transportation Networks.* Transportation Science 6 : 73–87.

[94] Dafermos S.C. (1980). *Traffic Equilibrium and Variational Inequalities.* Transportation Science 14 : 42–54.

[95] Dafermos S.C. (1982). *The general multimodal network equilibrium problem with elastic demand* Networks 12: 57-72.

[96] Daganzo C. (1983). *Stochastic Network equilibrium problem with multiple vehicle types and asiymmetric, indefinite link cost jacobians.* Transportation Science 17: 282-300.

[97] Daganzo C.F. (1979). *Multinomial probit: the theory and its application to demand forecasting.* Academic press, New York.

[98] Daganzo C.F. (1983). *Stochastic Network Equilibrium with Multiple Vehicle Types and Asymmetric, Indefinite Link Cost Jacobians.* Transportation Science 17: 282–300.

[99] Daganzo C.F., and M. Kusnic (1992). *Another look at the nested logit model.* Technical Report UCB-ITS-RTR 92-2, Institute of Transportation Studies, University of California, Berkeley.

[100] Daganzo C.F., and Y. Sheffi (1977) *On stochastic models of traffic assignment* Transportation Science 11: 253-274.

[101] Daganzo C.F., and Y. Sheffi (1982). *Unconstrained Extremal formulation of Some Transportation Equilibrium Problems.* Transportation Science 16: 332–360.

[102] Daly A.J., and S. Zachary (1978), *Improved multiple choice models.* in D.A. Hensher and M.Q. Dalvi (eds.), Determinants of Travel Choice. Saxon House, Westmead.

[103] Damberg O., J.T. Lundgren, and M. Patriksson (1996) *An algorithm for the stochastic user equilibrium problem* Transportation Research 30B : 115-131.

[104] Davis G.A. (1994). *Exact local solution of the continuous network design problem via Stochastic User Equilibrium assignment.* Transportation Research 28 B: 61-75.

[105] Davis G.A. N.L. Nihan (1993). *Large population approximations of a general stochastic traffic assignment model.* Operations Research 41: 169-178.

[106] Di Gangi M. (1988). *Una valutazione delle prestazioni statistiche degli estimatori della matrice che combinano i risultati di indagini e/o modelli con i conteggi di flussi di traffico.* Ricerca Operativa 51: 23-59.

[107] Dial R.B. (1971). *A Probabilistic Multipath Traffic Assignment Model with Obviates Path Enumeration.* Transportation Research.

[108] Dial R.B. (1996). *Bicriterion Traffic Assignment: Basic Theory and Elementary Algorithms.* Transportation Science 30/2: 93-111.

[109] Domencich T. A., and D. McFadden (1975). *Urban travel demand: a behavioural analysis.* American Elsevier, New York.

[110] Drake, Shofer, and May (1967). *A statistical analysis of speed-density hypotheses*. Highway Research Record 154, Transportation Research Board.

[111] Environmental Protection Agency (EPA) (1996). *Travel Survey Manual*. Department of Transportation, Washington D.C.

[112] Ferrari P. (1995). *Road pricing and network equilibrium*. Transportation Research 29 B: 357-372.

[113] Ferrari P. (1996). *Appunti di Pianificazione dei Trasporti*. 2nd edition, Editorial service of the University of Pisa, Italy.

[114] Ferrari P. (1997). *The Meaning of Capacity Constraint Multipliers in The Theory of Road Network Equilibrium*. Rendiconti del circolo matematico di Palermo, Italy, 48: 107-120.

[115] Fisk C. (1980) *Some Developments In Equilibrium Traffic Assignment Methodology*. Transportation Research B: 243-255.

[116] Fisk C. (1984). *Game theory and transportation systems modeling*. Transportation Research 18 B: 301-313.

[117] Florian M, and Spiess H., (1982), *The convergence of diagonalization algorithms for asymetric network equilibrium problems*. Transportation Research 16B: 477-483.

[118] Florian M. and D. Hearn (1992). *Networks equilibrium models and algorithms*.

[119] Florian M., and H. Spiess (1982). *The convergence of diagonalization algorithms for asymmetric network equilibrium problems* Transportation Research 16B: 477-483.

[120] Florian M., M. Gaudry, and P. Lardinois (1988). *A two dimensional framework for the understanding of Transportation planning models*. Transportation Research 22 B : 411-19.

[121] Foulds L.R. (1981). *A multicommodity flow network design problem*. Transportation Research 15 B: 273-283.

[122] Friesz T. L., D. Bernstein, T.E. Smith, R. L. Tobin, and B.W. Wie (1993). *A variational inequality formulation of the dynamic network users equilibrium problem*. Operations Research 41:179-191.

[123] Friesz T.L. (1985). *Transportation network equilibrium design and aggregation: Key developments and research opportunities*. Transportation Research 10 A: 413-427.

[124] Friesz T.L., J. Luque, R.L. Tobin, and B.W. Wie (1989). *Dynamic network traffic assignment considered as continuous time optimal control problem*. Operations Research 37: 893-901.

[125] Friesz T.L., R.L. Tobin, and P.T. Harker (1983). *Predictive intercity freight network models: the state of the art*. Transportation Research 17A: 409-417.

[126] Fukushima M. (1984). *A modified Frank-Wolfe algorithm for solving the traffic assignment problem* Transportation Research 18B: 169-178.

[127] Gallo G. and S. Pallottino (1988). *Shortest path algorithms*. In Fortran Codes for network Optimization, edited by B. Simeone, P. Toth, G. Gallo, F. Maffioli, and S. Pallottino. Annals of Operations Research 13: 3-79.

[128] Gallo G., Longo G., Nguyen S. and Pallottino S. (1993). *Directed hypergraphs and applications.* Discrete Applied Mathematics 2: 177–201.
[129] Greenshields B. (1934). *A study of traffic capacity.* Proceedings of the Highway Research Board 14, Transportation Research Board.
[130] Gunn H., J. and J.J. Bates (1982). *Statistical aspects of travel demand modelling.* Transportation Research 16A: 371-382.
[131] Haimes Y. Y., and V. Chankong (1985). *Decision Making with multiple objectives.* Springer Verlag, Berlin.
[132] Harker P.T. (1985). *Spatial Price Equilibrium: Advances in Theory, Computation and Application.* Springer Verlag, Heidelberg.
[133] Harker P.T. (1987). *Predicting Intercity Freight Flows.* Topics in Transportation, VNU Science Press, Utrecht, The Netherlands.
[134] Hearn D.W. (1997). *Toll pricing models for traffic networks.* Preprints of the 8th IFAC/IFIP/IFORS Symposium, Chania, Greece.
[135] Hearn D.W., S. Lawphongpanich, and S. Nguyen (1984). *Convex Programming Formulation of the Asymmetric Traffic Assignment Problem.* Transportation Research 18B: 357–365.
[136] Hensher D.A., Barnard P.O., and Truong T.P. (1988), *The role of Stated Preference methods in studies of travel choice.* Journal of Transport Economics and Policy 22.
[137] Heydecker B. G., and T. K. Khoo (1990). *The equilibrium network design problem.* Proceedings of AIRO '90, Sorrento, Italy: 587-602.
[138] Hickman M.D., and D.H. Bernstein (1997). *Transit service and path choice models in stochastic and time-dependent networks.* Transportation Science 31: 129-146.
[139] Hickman M.D., and N.H.M. Wilson (1995). *Passenger travel time and path choice implications of real-time transit information.* Transportation Research 3C: 211-226.
[140] Horowitz J.L. (1981). *Identification and diagnosis of specification errors in the multinominal logit model.* Transportation Research 15B: 345-360.
[141] Horowitz J.L. (1982). *Air quality analysis for urban transportation planning.* MIT Press, Cambridge, Mass.
[142] Horowitz J.L. (1982). *Evaluation of usefulness of two standard goodness of fit indicators for comparing non-nested random utility models.* Transportation Research Records 874.
[143] Horowitz J.L. (1984). *The stability of stochastic equilibrium in a two link transportation network* Transportation Research 18B: 13-28.
[144] Horowitz J.L. (1985). *Travel and location behaviour: state of the art and research opportunities.* Transportation Research 19A: 441-454.
[145] Horowitz J.L., J.M. Spermann, and C.F. Daganzo (1982). *An investigation on the accuracy of the Clark approximation for the multinomial probit model.* Transportation Science 16: 382-401.
[146] Hurdle V.F. (1984). *Signalized intersection delay models: a primer for the uniniziated.* Transportation Research Record 971.

[147] Hutchinson G. (1974). *Principles of urban transportation systems planning.* McGraw-Hill, New York.

[148] Institute of Transportation Engineers (1982). *Transportation and Traffic Engineering Handbook.* Prentice-Hall, Englewood Cliffs.

[149] Isard W. (1951). *Interregional and regional input-output analysis: a model of a space-economy.* The Review of Economics and Statistics 33: 318-328.

[150] Janson B. N. (1989). *Dynamic traffic assignment for urban road network.* Transportation Research 25B: 143-161.

[151] Jara-Diaz S. R. , and T. L. Friesz (1982). *Measuring the benefits derived for a transportation investment.* Transportation Research 16B: 57-77.

[152] Jayakarishnan R., H. Mahamassani, and T. Hu (1994). *An evaluation tool for advanced traffic information and management system in urban networks.* Transportation Research 2C:129-147.

[153] Jha M., S. Madanat, and S. Peeta (1998). *Perception updating and day-to-day travel choice dynamics in traffic networks with information provision.* Transportation Research 6C:189-212.

[154] Joliffe J.K., and T.P. Hutchinson (1975). *A behavioural explanation of the association between bus and passenger arrivals at a bus stop.* Transportation Science: 248-281.

[155] Kimber R.M., and E.M. Hollis (1978). *Peak period delays at road junctions and other bottlenecks.* Traffic Engineering and Control 19.

[156] Kimber R.M., M. Marlow, and E.M. Hollis (1977). *Flow-delay relationships at major-minor junctions.* Traffic Engineering and Control 18.

[157] Kitamura R., L. Kostyniuc, and K.L. Ting (1979). *Aggregation in Spatial Choice Modelling.* Transportation Science 13: 325-342.

[158] Kleinrock L. (1975). *Queuing System.* Vol. I and II, Jon Wiley & Sons, New York.

[159] Koppelman F.S. (1989). *Multidimensional Model System for Intercity Travel Choice Behaviour.* Transportation Research Records 1241: 1-8.

[160] Koppelman F.S., and J. Hauser (1978). *Destination Choice Behavior for Non-Grocery-Shopping Trips.* Transportation Research Records 673: 157-165.

[161] Langdon M.G. (1984). *Improved algorithms for estimating choice probabilities in the multinomial probit model.* Transportation Science 18: 267-299.

[162] Le Blanc L. J. (1975). *An algorithm for the discrete network design problem.* Transportation Science 9: 183-199.

[163] Le Blanc L.J. (1988). *Transit system network design.* Transportation Research 22 B: 383-390.

[164] Le Blanc L.J., and D.E. Boyce (1986). *A bilevel programming algorithm for exact solution of the network design problem with user optimal flows.* Transportation Research 20 B: 259-265.

[165] Le Blanc L.J., E.K. Morlook and W.P. Pierskalla (1975). *An efficient approach to solving the road network equilibrium traffic assignment problem* Transportation Research 9 : 309-318.

[166] Leontief W., and P. Costa (1987). *Il trasporto merci e l'economia italiana. Scenari di interazione al 2000 e al 2015.* Sistemi Operativi, New York-Venice.

[167] Leurent F. (1993). *Cost versus Time Equilibrium over a Network.* Eur. J. of Operations Research 71: 205-221.

[168] Leurent, F. (1995). *The practice of a dual criteria assignment model with continuously distributed values-of-time.* In Proceedings of the 23rd European Transport Forum: Transportation Planning Methods, E, 117-128 PTRC, London.

[169] Leurent, F. (1996). *The Theory and Practice of a Dual Criteria Assignment Model with a Continuously Distributed Value-of-Time.* In Transportation and Traffic Theory, J.B. Lesort ed., 455-477, Pergamon, Exeter, England.

[170] Louviere J.J. (1988), *Conjoint analysis modelling of Stated Preferences.* Journal of Transport Economics and Policy 22.

[171] Lupi M. (1986) *Convergence of the Franke-Wolf algorithm in transportation networks* Civ. Enging. Syst. 3.

[172] Lupi M. (1996). *Determinazione del livello di servizio di una intersezione urbana: alcune osservazioni sulla modellizzazione del ritardo.* Proceedings of "V Convegno SIDT", Naples.

[173] M. Florian, and Spiess H. (1989). *Optimal strategies: a new assignment model for transit networks.* Transportation Research 23B: 82-102.

[174] Magnanti T.L., and R.T. Wong (1984). *Network design and transportation planning: models and algorithms.* Transportation Science 18: 1-55.

[175] Mahamassani H., and Y. H. Liu (1999). *Dynamics of commuting decision behaviour under Advanced Traveller Information Systems.* Transportation Research 7C: 91-107.

[176] Maher M.J., and P.C. Hughes (1997*). A Probit-Based Stochastic User Equilibrium Assignment Model.* Transportation Research 31B: 341-355.

[177] Maher. M. J. (1983). *Inferences on trip matrices from observation on link volumes: a statistical approach.* Transportation Research 7B: 435-447.

[178] Manheim M. (1979). *Fundamentals of transportation systems analysis.* MIT Press, Cambridge, Mass.

[179] Manski C. (1977). *The structure of random utility models.* Theory and Decision 8: 229-254.

[180] Manski C.F., and D. McFadden (1981). *Alternative estimators and sample designs for discrete choice analysis.* in Structural Analysis of discrete data with Econometric Applications, MIT Press, Cambridge, Mass.

[181] Manski C.F., and Lerman S.R. (1977). *The estimation of probabilities from choice-based samples.* Econometrica 45.

[182] Marcotte P. (1983). *Network optimization with continuous control parameters.* Transportation Science 17: 181-197.

[183] Marcotte P. and D. Zhu (1996*). An Efficient Algorithm for a Bicriterion Traffic Assignment Problem.* In Advanced Methods in Transportation Analysis, L.Bianco, P.Toth eds., 63-73, Springer-Verlag, Berlin.

[184] Marcotte P., S. Nguyen and K. Tanguay (1996). *Implementation of an Efficient Algorithm for the Multiclass Traffic Assignment Problem.* In Transportation and Traffic Theory, J.B. Lesort ed., 455-477, Pergamon, Exeter, England.

[185] May A.D. (1990). *Traffic Flow Fundamentals.* Prentice Hall, Englewood Cliffs.

[186] McFadden, and D. (1978). *Modeling the choice of residential location.* in A.K. et al. (ed.), Spatial interaction theory and residential location: 75-96, North-Holland, Amsterdam.

[187] McShane W.R., and R.P. Roess (1990). *Traffic Engineering.* Prentice Hall, Englewood Cliffs.

[188] Merchant D. K., and Nemhauser G.L. (1978). *A model and an algorithm for the dynamic traffic assignment problem.* Transportation Science 12:183-199.

[189] Meyer M. D., and E. J. Miller (1984). *Urban Transportation Planning.* McGraw-Hill, New York.

[190] Mirchandani P., and H. Soroush (1987). *Generalized Traffic Equilibrium with Probabilistic Travel Times and Perceptions.* Transportation Science 21: 133-152.

[191] Mishan E. J. (1974). *Analisi Costi-Benefici.* Etas Books, Milan.

[192] Modenese Vieira L.F. (1992). *The value of service in freight transportation.* Ph.D. thesis, MIT, Boston.

[193] Montella B., and E. Cascetta (1978). *Tempo di attesa alle fermate di un servizio di trasporto collettivo urbano.* Ingegneria Ferroviaria 9: 827-832.

[194] Montella B., and M. Gallo (1998). *Metodologie per la progettazione delle reti di trasporto e loro applicazione: l'utilizzo di MT-Model.* Sistemi di Trasporto 3-4: 30-45.

[195] Montella B., M. Gallo, and R. Amirante (1998). *A general bus network design model and its application.* Preprints of TRISTAN III - Puerto Rico - 17-23 June 1998.

[196] Newell G.F. (1971). *Application of queuing theory.* Chapman and Hall LTD, London.

[197] Newell G.F. (1980). *Traffic flows in transportation networks.* MIT Press, Cambridge, Mass.

[198] Nguyen S. (1976). *A Unified Approach to Equilibrium Methods for Traffic Assignments.* In Traffic Equilibrium Methods, M. Florian editor, vol.118 of Lecture Notes in Economics and Mathematical Systems: 148–182.

[199] Nguyen S. and S. Pallottino (1988). *Equilibrium Traffic Assignment for Large Scale Transit Networks.* European Journal of Operational Research 37: 176–186.

[200] Nguyen S., and C. Dupuis (1984). *An efficient method for computing traffic equilibria in a network with asymmetric transportation costs* Transportation Science 18: 185-202.

[201] Nguyen S., and S. Pallottino (1986). *Assegnazione dei passeggeri ad un sistema di linee urbane: determinazione degli ipercammini minimi.* Ricerca Operativa 39: 207-230.

[202] Nguyen S., Morello E., and Pallottino S. (1989). *Discrete time dynamic estimation model for passenger origin-destination matrices on transit networks.* Transportation Research 22B: 251-260.

[203] Nguyen S., S. Pallottino, and M. Gendreau (1993*). Implicit Enumeration of Hyperpaths in a Logit Model for Transit Networks* Publication CRT 84.

[204] Nielsen O. A. (1997). *On The Distributions Of The Stochastic Components In SUE Traffic Assignment Models*, In Proceedings of 25th European Transport Forum Annual Meeting, Seminar F On Transportation Planning Methods, Volume II.

[205] Nijkamp P., P. Rietveld, and H. Voogd (1990). *Multicriteria Evaluation in Physical Planning.* North Holland Publication, Amsterdam, Netherlands.

[206] Nuzzolo A., and F. Russo (1995). *A disaggregate freight modal choice model.* Proceedings of 7th WCTR, Sydney, Australia.

[207] Nuzzolo A., and F. Russo (1996). *Stochastic assignment models for transit low frequency services: some theoretical and operative aspects.* In L. Bianco and P. Toth (editors), "Advanced methods in transportation analysis". Springer Verlang, New York, USA.

[208] Nuzzolo A., and F. Russo (1997). *Modelli per l'analisi e la simulazione dei sistemi di trasporto collettivo.* Ed. Franco Angeli, Rome.

[209] Nuzzolo A., and Russo F. (1993). *Un modello di rete diacronica per l'assegnazione dinamica al trasporto collettivo extraurbano.* Ricerca Operativa 67: 37-56.

[210] Nuzzolo A., F. Russo, and U. Crisalli (1999). *A doubly dynamic assignment model for congested urban transit networks.* Proceedings of 27th European Transportation Forum, Cambridge, England, Seminar F: 185-196.

[211] Nuzzolo A., U. Crisalli, and F. Gangemi (2000). *A behavioural choice model for the evaluation of railway supply and pricing policies.* Transportation Research 35A: 211-226.

[212] Oi W.Y., and P.W. Shuldiner (1972). *An Analysis of Urban Travel Demands.* Northwestern University Press, Evanston.

[213] Okutani I., Stephanades Y. (1984). *Dynamic prediction of traffic volume through Kalman Filtering theory.* Transportation Research 18B: 1-11.

[214] Oppenheim N. (1994). *Urban Travel Demand Modelling.* J. Wiley, New York.

[215] Ortuzar J. De D. (1992), *Stated Preference in travel demand modelling.* 6th World Conference on Transportation Research, Lyon.

[216] Ortuzar J.de D., and L.G. Willumsen (1994). *Modelling Transport.* John Wiley and Sons, 2nd edition.

[217] Papageorgiou M. ed. (1991). *Concise Encyclopedia of Traffic & Transportation System.* Pergamon Press.

[218] Papola A. (1996). *I modelli di Valore Estremo Generalizzato (GEV) per la simulazione della domanda di trasporto.* Internal report. Department of Transportation Engineering, University of Naples "Federico II".

[219] Papola A. (2000). *Some developments on the cross-nested logit model.* Proceedings of the IX IATBR conference, Gold Coast, Queensland, Australia.

[220] Patriksson M. (1994*). The Traffic Assignment Problem: Model and Methods.* VSP, Utrecht, The Netherlands.

[221] Payne H.J. (1971). *Models of freeway traffic and control.* Simulation Council Proc. 1: 51-61.

[222] Pearmin D., Swanson J., Kroes E., and Bradley M. (1991). *Stated Preferences Techniques: a guide to practice.* Steer Davies Gleave and Hague Consulting Group, London.

[223] Pignataro L.J. (1973). *Traffic engineering, theory and practice.* Prentice-Hall, Englewood Cliffs.

[224] Poorzahedy H., and M. A. Turnquist (1982). *Approximate algorithms for the discrete network design problem.* Transportation Research 16 B: 45-55.

[225] Potts R.B., and R.M. Oliver (1972). *Flows in transportation networks.* Academic Press, New York.

[226] Powell W.B. and Y. Sheffi (1982). *The Convergence of Equilibrium Algorithms with Predetermined Step Sizes.* Transportation Science 16: 45–55.

[227] Ran B., and D. E. Boyce (1994). *Dynamic urban transportation network models: theory and implications for Intelligent Vehicle-Highway Systems.* Springer-Verlag.

[228] Richards M., and M. Ben Akiva (1975). *A Disaggregate Travel Demand Model.* D.C. Heath, Lexington, Mass.

[229] Road Research Laboratory (RRL) (1965). *Research on road traffic.* RRL, London.

[230] Robertson D.I. (1979). *Traffic models and optimum strategies of control: a review.* Proceedings of the International Symposium on Traffic Control Systems. Edited by W.S. Homburger and L. Steinman, vol. 1, University of California, Berkeley, California: 262-288.

[231] Russo F., and A. Vitetta (1995). *Networks and assignment models for the Italian national transportation systems.* Proceeding of the 7th WCTR, Sidney, Australia.

[232] Russo F., and A. Vitetta (1998*). A C-Logit assignment without explicit path enumeration* Quaderno del dipartimento di Informatica, Matematica, Elettronica e Trasporti, University of Reggio Calabria.

[233] Seddon P.A., and M.P. Day (1974). *Bus passenger waiting time in Greater Manchester.* Traffic Engineering and Control: 442-445.

[234] Sheffi Y. (1985). *Urban transportation networks.* Prentice Hall, Englewood Cliff, NJ.

[235] Sheffi Y., and W.B. Powell (1982), *An Algorithm For The Equilibrium Assignment Problem With Random Link Times*, Networks 12.

[236] Sheffi Y., and W.B. Powell (1983). *Optimal signal settings over transportation networks.* Journal of Transportation Engineering 109: 824-839.

[237] Small K.A. (1982). *The scheduling of commuter Activities: work trips.* The American Economic Review 72: 467-479.

[238] Small, and K. (1987). *A discrete choice model for ordered alternatives.* Econometrica 55: 409-424.

[239] Smith M.J. (1979*). The Existence, Uniqueness and Stability of Traffic Equilibrium.* Transportation Research 13B: 295–304.

[240] Stopher P. R., and A. H. Meybourg (1976). *Transportation systems evaluation.* Lexington Books, Lexington, Mass.

[241] Thomas R. (1991). *Traffic Assignment Techniques.* Avebury Technical, England.

[242] Transportation Research Board (1997). *Highway Capacity Manual.* Special Report 209, Washington D.C., 3th edition.

[243] Van Vliet D. (1981*) Selected Node-Pair Analysis in Dial's Assignment Algorithms* Transportation Research 15B: 65-68.

[244] Van Vliet D. (1987) *The Frank-Wolfe Algorithm for Equilibrium Traffic Assignment Viewed as a Variational Inequality* Transportation Research 21: 87-89.

[245] Van Zuylen J.H., and Willumsen L.G. (1980). *The most likely trip matrix estimated from traffic counts.* Transportation Research 14B: 281-293.

[246] Voogd H. (1983). *Multicriteria Evaluation for Urban and Regional Planning.* Pion Ltd, London, Great Britain.

[247] Vovsha P. (1997). *The Cross-Nested Logit Model: Application to Mode Choice in the Tel-Aviv Metropolitan Area.* Proceedings of the 76th TRB Meeting.

[248] Vovsha P. S. Bekor (1998). *The link-Nested Logit Model of Route Choice: Overcoming the Route Overlapping Problem.* Proceedings of the 77th TRB Meeting.

[249] Vytoulkas P.K. (1990). *A dynamic stochastic assignment model for the analysis of general networks.* Transportation Research 24B: 453-469.

[250] Wachs M. (1985). *Planning, organizations and decision-making: a research agenda.* Transportation Research 19A: 521-532.

[251] Wardrop J.G. (1952). *Some Theoretical Aspects of Road Traffic Research.* Proc. Inst. Civ. Eng. 2: 325–378.

[252] Watanatada T., and M. Ben Akiva (1979). *Forecasting urban travel demand for quick policy analysis with disaggregate choice model: a Monte-Carlo simulation approach.* Transportation Research 13A: 241-248.

[253] Watling D.P. (1996). *Asymmetric problems and stochastic process models of traffic assignment.* Transportation Research 30B: 339-357.

[254] Watling D.P. (1999). *Stability of the stochastic equilibrium assignment problem: a dynamical systems approach.* Transportation Research 33B: 281-312.

[255] Webster F.V. (1958). *Traffic signal settings.* Road Research Technical Paper 39, HMSO, London.

[256] Webster F.V. e Cobbe B.M. (1966). *Traffic Signals.* Road Research Technical Paper 56, HMSO, London.

[257] Wie B. W., T. L. Friesz, and T. L. Tobin (1990). *Dynamic user optimal traffic assignment on congested multi-destination networks.* Transportation Research 24B: 431-442.

[258] Williams H.C.W.L. (1977). *On the formation of travel demand models and economic evaluation measures of user benefit.* Environment and Planning A: 285-344.

[259] Williams H.C.W.L., and J. de D. Ortùzar (1982). *Behavioural theories of dispersion and the mis-specification of travel demand models.* Transportation Research 16B: 167-219.

[260] Willumsen L.G. and Tamin O.Z. (1989) *Transport demand model estimation from traffic counts.* Transportation 16: 3-26.

[261] Wilson A.G. (1974). *Urban and regional models in geography and planning.* John Wiley & Sons, London.

[262] Winston C. (1983). *The demand for Freight Transportation: Models and Applications.* Transportation Research 17A: 419-427.

[263] Wohl M., and B. V. Martin (1967). *Traffic system analysis for engineers and planners.* McGraw-Hill, New York.

[264] Wu J. H., and M. Florian (1993). *A Simplicial Decomposition Method for the Transit Equilibrium Assignment Problem,* Annals of Operations Research.

[265] Wu J.H., M. Florian, and P. Marcotte (1994). *Transit Equilibrium Assignment: a Model and Solution Algorithms.* Transportation Science 28: 193–203.

[266] Wu.J. H., Y. Chen, and M. Florian (1995). *The continuous dynamic network loading problem: a mathematical formulation and solution method.* Transportation Research 32B: 173-187.

[267] Xu Y., J. Wu, M. Florian, P. Marcotte, and D.L. Zhu (1994). *New advances in the continuous dynamic network loading problem.* Forthcoming in Transportation Science.

[268] Yang H., and S. Yagar (1995). *Traffic assignment and signal control in saturated road networks.* Transportation Research 29 A: 125-139.

[269] Yang Q., and H. Koutsopoulos (1996). *A microscopic traffic simulator for evaluation of dynamic traffic management systems.* Transportation Research 4C: 113-129.

[270] Yates F. (1981). *Sampling methods for censuses and surveys.* Griffin, London.

Index

Access/egress time 63
Accessibility 4, 185
  active _ 4
  passive _ 4
  activity system 2
  _ attributes 99
Aggregation method(s) 151
  average individual _ 152
  classification _ 153
  sample enumeration _ 153
  sample enumeration and classification _ 154
  target variables _ 155
Algorithm(s)
  _ with ordering (Label-Setting) 438
  _ without ordering (Label-Correcting) 438
  assignment _ *see* Assignment algorithm(s)
  bisection _ 664
  conjugate gradient _ 668
  diagonalization _ 464
  Dial _ 441
  Frank-Wolfe _ 670
  golden section _ 662
  gradient _ 666
  projected gradient _ 669
  shortest hyperpath _ 466
  shortest path _ 436
  solution _ for fixed point problems 656
  solution _ for optimization problems 660
All-Or-Nothing algorithm 453
Alternative Specific Attribute (ASA) 99, 104, 193
Alternative Specific Constant (ASC) 99, 104, 193
Analysis interval 12
Analysis system 1
Arc(s) *see* Link(s)
Arrival(s)
  _ curve 80
  _ pattern 85
Assignment algorithm(s)
  algorithms for assignment with pre-trip/en-route path choice 466
  algorithms for rigid demand User Equilibrium (UE) assignment 456
  algorithms for rigid demand User Equilibrium (UE) assignment with pre-trip/en-route path choice 472
  algorithms for System Optimal (SO) assignment 465
  algorithms for Uncongested Network (UN) assignment 440
  algorithms for Uncongested Network (UN) assignment with pre-trip/en-route path choice 471
  All-Or-Nothing (AON) _ 453
  Dial algorithm 441
  explicit path enumeration 440
Assignment models 19, 251
  _ algorithms *see* Assignment algorithm(s)
  _ with pre-trip/en-route path choice 297
  Deterministic Uncongested Network (DUN) _ 271, 451, 453
  Deterministic Uncongested Network (DUN) assignment map 272
  Deterministic User Equilibrium (DUE) _ 281, 282, 307

Deterministic User Equilibrium (DUE) properties 281
Dynamic Network Loading (DNL) 384, 395
Dynamic Traffic _ (DTA) *see* Dynamic
elastic demand User Equilibrium (UE) _ 307, 310
inter-period Dynamic Process assignment models 331
multi-class _ 324
multi-mode _ 321
optimization models for stochastic assignment 357
rigid demand _ 255
rigid demand Deterministic User Equilibrium (DUE) 460
rigid demand Stochastic User Equilibrium (SUE) 457
rigid demand User Equilibrium (UE) _ 274
single-mode _ 255, 309, 311
single-user class _ 255
Stochastic Uncongested Network (SUN) _ 270, 327, 357, 441, 450
_ _ with Logit path choice model 441
_ _ without explicit paths enumeration 441
_ _ with Probit path choice model 448
Stochastic User Equilibrium (SUE) _ 276, 357
_ assignment function 270
_ properties 278
System Optimal _ 291
Uncongested Network (UN) _ 268, 440, 462, 473, 485
undifferentiated congestion multi-class _ 328
within-day Dynamic Traffic Assignment (DTA) models 403

ATIS 413
Attractive lines 62, 209
Attribute(s) 96, 98
activity system _ 99
Alternative Specific _ (ASA) 99, 104, 193
attractivity _ 191
cost _ 191
generic _ 99
Level Of Service (LOS) _ 99, 193
performance _ 26, 99
socio-economic _ 99, 186, 194
specific _ 99
Availability/perception variable 193
Average service time 83

Backward
_ shortest paths 436
_ star 444
_ travel time 381
_ tree 438
Base network 8
Bayesian estimator(s) *see* Estimator(s)
Behavioral models *see* Transportation demand model(s)
Bellman principle 437
Benefit-Cost analysis 625
Box-Cox transformation 99
BPR *see* Cost function(s)
Braess paradox 293
Brouwer's theorem 282, 654

Calibration
_ of demand models 494, 517
aggregate _ of demand models using traffic counts 542
Capacity 44, 70
urban road network _ 578
Car ownership model 217
Centroid(es) *see* Node(s)

Choice(s)
_ alternative(s) 95, 96
_ dimension(s) 95, 175
_ model(s) *see* Transportation demand model(s)
_ probability 96
_ updating model 334
mobility _ 176
travel _ 176
Choice set 96, 98, 193, 198
_ modeling 137
Choice tree
_ of Multi-Level Hierarchical Logit 113
_ of Multinomial Logit 102
_ of Single-Level Hierarchical Logit 107, 110
Cholesky factorization 133
Clark approximation 134
C-Logit *see* Logit model(s)
Commonality factor 205
Conditional path choice matrix 299
Congested system 69
Congestion 3, 34, 69
differentiated _ 326
undifferentiated _ 326
Connector *see* Link(s)
Conservation differential equations 76
Continuous service systems 39
Cordon centroid 37
Cost(s)
_ attributes *see* Attribute(s)
actual path costs 410
additive _ 214
additive and class specific _ 329
additive and generic _ 329
additive path _ 30
expected path _ 410
generalized transportation _ 30, 43
generalized transportation link _ 26
hyperpath additive _ 298
hyperpath non additive _ 298
link _ function 34
link _ vector 28
link-wise additive path _ 256
link-wise non-additive path _ 256
non additive _ 214
non additive path _ 30
path _ vector 30
reference _ 328
transportation _ 26
Cost function(s) 33, 43
_ for scheduled systems 61
_ vector 35
Ackcelik _ 53
BPR _ 44, 73
extra-urban _ 45
hyperbolic _ 73
intersection _ 49
link _ 34
parking link _ 58
polynomial _ 73
toll-barrier _ 46
urban road _ 48
Webster _ 51, 91
Critical density 70
Critical speed 70
Cross elasticity *see* Elasticity
Cumulated frequency 62
Cycle length 50, 87

Decision-maker (rational) 96
Delay 51, 78
_ for signalized intersections 51, 87
deterministic _ functions 84
deterministic _ models 90
queuing _ *see* Queuing
schedule _ *see* Schedule delay
stochastic _ models 91
total _ 83
total _ models 91

Demand
_ estimation *see* Estimation
_ model(s) *see* Transportation demand model(s)
see Logit model(s)
see Probit model(s)
_ temporal variation *see* Temporal variations of Transportation Demand
elastic _ 307, 314, 318
_ _ equilibrium assignment models 307
_ _ single-mode Deterministic User Equilibrium models 316, 319
_ _ single-mode Stochastic User Equilibrium models 313, 314
intermediate _ 234
Density 67
jam _ 70
Departure curve 80
Destination choice model *see* Distribution model
Deterministic choice model 98
_ (mathematical properties) 145
Deterministic process 336
_ models 339
Diagonalization algorithm 465
Dial algorithm 441
Distribution model(s) 179, 188, 218
Diversion probability 210
Driving license holding model 215
Dynamic(s)
_ demand-supply interaction models 403, 406, 410
_ supply models 370, 388
_ user equilibrium 406
day-to-day _ 331
within-day _ 403
Dynamic Network Loading (DNL)
continuous models 384
discrete models 395
Dynamic process 336, 410
_ algorithm 344
_ modeling 336
deterministic _ 413
stochastic _ 413

Economic
_ activity sectors 230, 233
_ analysis 606
_ impacts 608
Elastic demand *see* Demand
Elasticity
_ of Hierarchical Logit models 149
_ of Logit models 148
_ of random utility models 147
cross _ 147
direct _ 147
link _ 148
point _ 148
Elastic trade coefficients 240
ELECTRE methods 636
Elementary destination 188
Emission model(s) 184
category index _ 185
category regression _ 186
En route choice behavior 209
Environmental impacts 609
Ergodic 346
Estimation
_ of demand models 485, 492
_ of demand variations (forecasting) 557
_ of intra-periodal dynamic demand flows using traffic counts 548
_ of O-D demand flows using traffic counts 522
_ of transportation demand 485, 555
direct _ of present demand 486
disaggregate _ 492, 508
dynamic _ of O-D demand flows 553

Estimator(s)
Bayesian _ 533
Generalized Least Squares (GLS) _ 532
Maximum Likelihood (ML) _ 529
Non Linear Generalized Least Squares (NLGLS) _ 545
sampling _ 488
sequential _ 554
simultaneous _ 553
Exchange trips 10
Expected Maximum Perceived Utility (EMPU) 103, 141, 613, 623
_ in Logit models 103
_ in Multi-Level Hierarchical Logit models 108, 114
Explicit path enumeration *see* Path choice
Feasibility studies 605
Feasible path and link flow sets 264

Factorial Design
fractional _ 515
full _ 512
FIFO rule 85, 376
condition for _ 377
Financial analysis 606
Fixed point problems 339, 652
_ states 340
properties of _ 654
solution algorithms for _ 656
stability regions of a _ state 344
Flow(s) 31, 67
_ conservation 69, 373
attracted _ 10
emitted _ 10
equivalent _ 31
link _ 31
link _ vector 31
path _ 33
path _ vector 33
relationship between link _ and path _ 259
relationship between stochastic and determistic equilibrium _ 288
saturation _ 50
user _ 31
vehicle _ 31
Fluid approximation 75
Free-flow speed 70
Freight transport demand model 230
Frequency 61
cumulated _ 62
Fundamental diagram of traffic flow 70

Game theory 566
Nash game 566
Stackelberg game 566
General flow conservation equation 66
Generalized Extreme Value (GEV) 126, 157
Goods typology 230
Graph 25
_ model(s) 25, 39
access _ 59
diachronic _ 59
line _ 59
line-based graph models 59
run _ 59
transportation _ 37
Gravitational model 192
Greenberg's model 72
Greenshield's model 72
Gumbel random variable 101, 165

Headway(s) 66
Hyperpath(s) 210
_ additive cost(s) 298
_ choice map 301
_ choice probability 301
_ non additive cost(s) 298

composed _ 210
shortest _ algorithms 467
simple _ 210
Hyper-network algorithms 480

Impact(s) 35
_ function 35
_ indicators 609
social _ 609
territorial _ 609
Implicit Availability Perception (IAP) models 139
Implicit path enumeration *see* Path choice
Inclusive variables *see* Expected Maximum Perceived Utility (EMPU)
Independence from Irrelevant Alternatives (IIA) 105
Industrial logistic characteristics 230
Inflow 67, 372
cumulative continuous _ 372
cumulative discrete _ 391
Information
descriptive vs prescriptive _ 413
historical _ 413
predictive _ 413
pre-trip vs en-route 413
real-time 413
Internal Return Rate (IRR) 629
Internal trips 11
Intersection(s)
_ link 49
priority _ 57
signalized _ *see* Signalized intersection(s)
Interzonal trips 6
Intra-period dynamics 13, 549
Intra-period stationarity 13
Intra-zonal trips 6
Inverse travel time function 375

Jam density 70
Journey 10
_ demand models *see* Trip chaining models
_ frequency model 224
_ type choice model 226

Lane group 50
Leaving-time function 375
Level Of Service (LOS) attributes *see* Attributes
Likelihood 494
Line-based graph models 59
Link(s) 25, 38
_ capacity 44
_ cost function 33
_ cost vector 28
_ elasticity 148
_ flow 31
_ flow conservation 69
_ flow vector 31
_ performance 33
_ performance function 34
_ performance models 379
exit function 379
travel time function 379
access _ 59
alighting _ 59
boarding _ 59
connector _ 38
dwelling _ 59
extra-urban road _ 45
intersection _ 49
line _ 59
motorway _ 44
on-board _ 60
parking _ 58
queuing _ *see* Queuing
real _ 38
running _ 65, 370
toll-barrier _ 46
urban road _ 48
waiting _ 59
Link-path incident matrix 25, 257

Logit model(s) 101, 157, 185, 202
C-Logit 205
Cross-Nested _ 122, 163
Hierarchical _ 106, 183, 194, 223
_ _ Multi-Level 113, 160
_ _ Single-Level 106, 158
Hybrid Logit-Probit model 136
Nested _ *see* Hierarchical _
Tree _ *see* Hierarchical Multi-Level _
Logsum variable *see* Expected Maximum Perceived Utility (EMPU)

Macroscopic models *see* Traffic flow(s)
Market segments 12
Markov process 345
Maximum Likelihood (ML) Method 494
Maximum Perceived Utility *see* Expected Maximum Perceived Utility (EMPU)
Mesoscopic models *see* Traffic flow(s)
Method of Successive Averages (MSA) 457, 657
_ Cost Averaging (MSA-CA) 457
_ Flow Averaging (MSA-FA) 457
Microscopic models *see* Traffic flow(s)
Mobility
_ choices 176
_ models 176
Modal preference 193
Modal split model(s) *see* Mode choice model(s)
Mode choice model(s) 179, 192, 219
consignment _ 244
logistic _ 244
Model(s) *see also* Transportation demand model(s)
_ calibration 494, 517
_ estimation 485, 492
_ specification 493, 557
_ validation 502
Monitoring 603
Monte Carlo method 133, 451
Motorway link 44
MSA *see* Method of Successive Averages (MSA)
Multi-Criteria analysis 625
Multinomial Logit model *see* Logit model(s)
Multi-objective optimization 571
Multi-Regional Input-Output (MRIO) models 232
_ with constant coefficients 236
_ with elastic prices 241
_ with elastic trade coefficients 238
_ with variable coefficients 238
matrix of technical coefficients 235
Multivariate Normal 129, 168

Nash game *see* Game theory
Neighborhood of finite radius 643
Nested Logit model *see* Logit model(s)
Net Present Value (NPV) 629
Network
_ Flow Propagation (NFP) 33, 258
_ layout variables 570
_ Loading (NL) 33
_ models 24
_ performance indicators 265
_ topology variables 567
base _ 8
diachronic _ 25, 419
synchronic _ 25
transportation _ 36

Node(s) 25
_ consistency equations 77
_ flow conservation 373
access _ 59
centroid _ 6, 25
cordon _ *see* Cordon centroid
diversion _ 59
line _ 59
stop _ 59
temporal centroid 419
zone centroid 6
Non-equilibrium models 331
Non-project state 611
Non-stationary models 75
Numerical analysis 643

O-D
_ flow 10
_ matrix 10
Offset *see* Signalized intersection(s)
On-board travel time 61
Opportunity costs 627
Optimization model(s)
_ for stochastic assignment 357
bi-level _ 570
Optimization problems 658
Out-flow 67, 372
cumulative continuous _ 372
cumulative discrete _ 391
Over-saturation 82

Packet(s) 388
Parking
_ choice model 179, 195
_ link 58
Partial share model 179
Path(s) 25
_ cost 30
_ cost vector 30
_ choice *see* Path choice
_ flow 33
_ flow vector 33
additive _ cost vector 30
backward shortest _ 436
efficient _ 441
explicit _ enumeration 440
feasible _ 264
forward shortest _ 436
link-wise additive _ costs 256
link-wise non-additive _ costs 256
non additive _ cost vector 31
shortest _ algorithms 436
Path choice
_ exhaustive approach 198, 213
_ model *see* Path choice model
_ selective approach 198, 213
explicit path enumeration 198
implicit path enumeration 198
Path choice model 179, 197, 219
_ for road systems 197
_ for transit systems 207
shortest hyperpath algorithms 468
shortest path algorithms 436
Performance
_ attribute *see* Attribute(s)
_ indicators 632
_ variable(s) 26, 28
_ vector 28
additive path _ variables 28
supply _ variables 570
Planning process 599
Point elasticity *see* Elasticity
Pre trip
_ choice 197
_ choice behavior 209
_/en route mixed choice 197
Price variables 567, 570
Primary activity destination 221
Priority intersection 57
Probability function 97, 127
Probit model(s) 128, 203
Factor-Analytic _ 130
Random Coefficients _ 131
Project state 611

Queuing
_ delay 83
_ link 65, 78, 370
_ theory 85
deterministic _ models 88
expected queue length 86
models for queuing links 78
stationary _ system 81
queue discipline 85
server 78

Random residuals 96, 260
Random utility model(s) 95, 215
_ for partial share 181
additive _ 97, 143
aggregation methods for _ 151
factorialization of _ 181
mathematical properties of _ 142
Reciprocal substitution
_ coefficients 100
_ rates 193
Renewal process 348
Revealed Preference (RP) 492, 508
Rigid demand 253, 275
_ Deterministic User Equilibrium 461
_ Stochastic User Equilibrium 457
Route choice model *see* Path choice model
Running time 48

Sample enumeration *see* Aggregation method(s)
Sampling
_ estimate 489
_ estimator(s) *see* Estimator(s)
_ rate 489
_ strategy 487
_ surveys *see* Survey(s)
_ unit 487
cluster _ 488
simple random _ 488, 494
stratified random _ 488, 490, 495
Saturation flow 50, 88
Schedule delay 399
early arrival penalty 399
early departure penalty 400
late arrival penalty 399
late departure penalty 400
Scheduled service(s)
_ system 58
_ with irregular high frequency 425
_ with regular low frequency 418
Selective approach 198, 213
Server *see* Queuing
Service line 58
Service pattern 85
Shadow prices 627
Shippers 231
Shortest
_ hyperpath algorithms 467
_ path algorithms 436
Signalized intersection(s) 49, 87
delay for _ *see* Delay
effective green/cycle ratio 50, 87
effective green time 50, 87
effective red time 87
traffic-ligth cycle 50, 87
Signal setting optimization *see* Supply Design Problem(s) (SDP)
Simulation model 568
Size
_ function 190
_ variable 190
Social objective function 570
Socio-economic attributes *see* Attribute(s)
Space continuous models 77
Space discrete models 76
Space mean speed 66
Spatial Price Equilibrium (SPE) models 231

Specification
_ of demand models 493
Specific attributes *see* Attribute(s)
Stackelberg game *see* Game theory
Stated Preferences (SP) 492, 508
Stationary flow conservation equation 69
Stationary models 68
Stationary probability distribution 346
Statistic(s)
_ and test on goodness of fit 506
Rho-square _ 506
Stochastic process 336, 348
_ models 345
Strategic planning 604
Study area 6
Supply Design model(s) *see* Supply Design Problem(s) (SDP)
Supply Design Problem(s) (SDP) 565
_ constraints 568
_ objectives 567
_ variables 567, 570
extra-urban road network capacity design 576
general supply design model 569
objective function 569
operator objective functions 571
optimal functional layout problem 574
optimal infrastructure layout problem 572
road network capacity design 576
road network layout design 572
traffic signal setting problem 576
transit network design 577
urban road network capacity design 576
Supply model 23, 36, 256
Supply performance variables 570
Surplus 613, 618
Survey(s)
_ design 493, 511
cordon _ 486
destination _ 487
household _ 487
mail _ 487
on-board _ 486
Revealed Preference (RP) _ 492, 508
sampling _ 486, 556
Stated Preference (SP) _ 492, 508
statistical design of sampling _ 487
while trip _ 486
Synchronic networks 25
System design phase 603
System Optimal assignment 291
System projects 601

Tactical planning 605
Target variable method *see* Aggregation method(s)
Technical coefficients 235
Temporal centroid 419
Temporal variations of Transportation Demand
cyclic variations 14
estimation of demand variations (forecasting) 557
long term variations (trends) 13
inter-period variations 14
Territorial impacts 609
Test(s)
_ on the functional form 507
Chi-square _ 504
formal _ on coefficients 503
informal _ on coefficient 502
Likelihood Ratio _ 505
T-student _ 504
Time mean speed 66
Time period choice model 179

Toll-barrier link 46
Topological variables 570
Total delay *see* Delay(s)
Trade
_ coefficients 237
_ matrix 237
Traffic assignment models *see* Assignment models
Traffic counts
aggregate calibration of demand models using _ 542
estimation of intra-periodal dynamic demand flows using _ 548
estimation of O-D demand flows using _ 522
Traffic flow(s) *see also* Flow(s)
_ theory 65
macroscopic _ models 75
mesoscopic _ models 75, 388
microscopic _ models 75, 388
Traffic-ligth *see* Signalized intersection(s)
Traffic signal setting problem *see* Supply Design Problem(s) SDP
Traffic zones 6
Transit line
attractive _ 62, 209
Transportation demand
_ level 179
_ model(s) *see* Transportation demand model(s)
temporal variations of _ *see* Temporal variations of Transportation Demand
Transportation demand model(s) 175, 215, 259, 555
activity participation _ 177
aggregate _ 177
aggregation procedure 220
behavioral models 177, 184, 188, 231
calibration of _ *see* Calibration
descriptive _ 177, 185, 231
deterministic utility model 202
disaggregate _ 177, 231, 244
elasticity of _ *see* Elasticity
emission model(s) *see* Emission model(s)
estimation of _ *see* Estimation
four-level (or four-stage) model 179
gravitational model 192
interpretative models *see* behavioral models
mobility _ 176
non interpretative models *see* descriptive models
parking choice model 179, 195
specification of _ *see* Specification
time period choice model 179
trip chaining models 176, 220
trip emission model *see* trip frequency model
trip frequency model 179, 184, 217
validation of _ *see* Validation
Transportation supply 3, 23
Transportation systems 1
Travel demand flow(s) 10
Travel models 176
Travel strategy 209
Tree Logit model *see* Logit model(s)
Triangolar inequality 436
Trip chaining models *see* Transportation demand model(s)
Trip demand model system *see* Transportation demand model(s)
Trip emission model *see* Emission model(s)
Trip frequency model *see* Emission model(s)

Under-saturation 81
Urban road
_ link 48
_ network capacity 576
User categories 178
User flows *see* Flow(s)
Util 100
Utility
_ updating model 333
perceived _ 96, 100
systematic _ 96, 98

Validation
_ of demand models 502
Value Of Time (VOT) 194, 502
Variance-Covariance matrix
_ of random residuals 98
_ of the Cross-Nested Logit model 122, 125
_ of the Factor Analytic Probit model 131
_ of the Hybrid Logit-Probit model 136
_ of the Logit model 102
_ of the Multi-Level Hierarchical Logit model 117
_ of the Probit model 129, 130
_ of the Random Coefficient Probit model 132
_ of the Single-Level Hierarchical Logit model 110
Variational inequality 281
_ problems 673
Vehicle flows *see* Flow(s)
Vertice(s) *see* Node(s)

Waiting
_ link *see* Link(s)
_ time 49, 61 *see also* Delay
Wardrop's
_ first principle 264, 272, 282
_ second principle 291
Webster formula 51, 91
Within-day 253

Zone centroids 6
Zoning 6

Main Variables

AL_m	set of attractive lines at diversion node m
AL_{mj}	set of attractive lines of hyperpath j at diversion node m
$\boldsymbol{b}_l$	vector of physical and functional characteristics of link l
$\boldsymbol{c}$	vector of link costs
$\boldsymbol{c}(\boldsymbol{f})$	vector of link cost functions
$\boldsymbol{c}^i$	vector of link costs for class i
c^i_l	generalized transportation cost of link l for user class i
c_l	generalized transportation cost of link l
$c_l(\boldsymbol{f})$	cost function for link l
$\boldsymbol{d}$	vector of demand flows
d_{od}	demand flow for the od pair
$d_{od,m}$	demand flow of the users between the pair od with mode m
$d_{od}[smkh]$	od demand flow for purpose s using mode m and path k during period h
$\boldsymbol{e}$	vector of impact variables
$\boldsymbol{e}(\boldsymbol{f})$	vector of impact functions
$e_l(f)$	impact function for link l
f_l	total flow on the link l
$\boldsymbol{f}$	vector of link flows
f^i_l	flow of user class i on link l
$\boldsymbol{f}^i$	vector of link flows for class i with entries f^i_l
$\boldsymbol{f}^{od}$	vector of link flows f^{od}_l
$\boldsymbol{f}_{DUN}$	vector of link flows in the deterministic uncongested network assignement
$\boldsymbol{f}_{SUN}$	vector of link flows in the stochastic uncongested network assignement
$\boldsymbol{f}_{UN}$	vector of link flows in the uncongested network assignment
g_k	generalized transportation cost of path k
$\boldsymbol{g}$	vector of total path costs
$\boldsymbol{g}(\boldsymbol{h})$	vector of path cost functions
g_k^{ADD}	additive cost of path k
$\boldsymbol{g}^{ADD}$	vector of additive path costs
g_k^{NA}	non additive cost of path k
$\boldsymbol{g}^{NA}$	vector of non additive path costs
$\boldsymbol{g}_{od}$	vector of total costs for paths connecting the od pair
$\boldsymbol{g}_{od}^{ADD}$	vector of additive costs g_k^{ADD} for paths connecting the od pair
$\boldsymbol{g}_{od}^{NA}$	vector of non additive costs g_k^{NA} for paths connecting the od pair
$\boldsymbol{g}_{od,m}$	vector of total path costs for the pair od and the mode m
$\boldsymbol{g}_{od,m}^{ADD}$	vector of additive path costs for the pair od and the mode m
$\boldsymbol{g}_{od,m}^{NA}$	vector of non additive path costs for the pair od and the mode m
h_k	flow on path k

$\boldsymbol{h}$	vector of path flows
$\boldsymbol{h}_{DUN}$	vector of path flows in the deterministic uncongested network assignment
$\boldsymbol{h}_{SUN}$	vector of path flows in the stochastic uncongested network assignment
$\boldsymbol{h}_{od}$	vector path flows of the users on the *od* pair
$\boldsymbol{h}_{od,i}$	vector path flows for the pair *od* and the class *i*
$\boldsymbol{h}_{od,m}$	vector path flows for the pair *od* and the mode *m*
I^i	choice set for decision maker *i*
(i,j)	oriented link between nodes *i* and *j*
j	hyperpaths index
$\boldsymbol{Jac}[\ldots]$	jacobian
J_{od}	set of hyperpaths connecting the *o,d* pair
$J_{od,m}$	set of hyperpaths connecting the *o,d* pair on the network of the transit mode *m*
k	path index
K_{od}	set of feasible routes connecting the centroid pair *o,d*
l	link index
mc	monetary cost
p_k or $p_{od,k}$	probability of choosing path *k*
$\boldsymbol{p}_{od}$	vector of path choice probabilities for the *od* pair
$\boldsymbol{p}_{od,i}$	vector of path choice probabilities for the *od* pair and the class *i*
$\boldsymbol{p}_{od,m}$	vector of path choice probabilities for the pair *od* and the mode *m*
$\boldsymbol{P}^{SIM}$	matrix of the path choice fractions resulting from the assignment model
$\boldsymbol{Q}$	matrix of hyperpath choice probabilities
Q	capacity
Q_l	capacity of link *l*
q_j *or* $q_{od,j}$	probability of choosing hyperpath *j*
$\boldsymbol{q}$ *or* $\boldsymbol{q}_{od}$	vector of probabilities for all hyperpaths between the same *o,d* pair
r_{nl}	performance attribute *n* for link *l*
$r_{nl}(\boldsymbol{f}; \boldsymbol{b}_{nl}, \gamma_{nl})$	performance attribute function *n* for link *l*
$\boldsymbol{r}_n$	vector of performance attributes r_{nl}
s	Expected Maximum Perceived Utility (EMPU)
$\boldsymbol{s}$	vector of path choice EMPUs
S	saturation flow
S_0	ideal saturation flow per lane
$\boldsymbol{SE}$	vector of socio-economic variables
Ta_j	total access/egress time in hyperpath *j*
Tal_j	total alighting time in hyperpath *j*
Tb_j	total on board time in hyperpath *j*
Tbr_j	total boarding time in hyperpath *j*
Td_j	total dwelling time in hyperpath *j*
Tw_j	total waiting time in hyperpath *j*

ta_l	access/egress time on link l
tal_l	alighting time on link l
tb_l	on board time on link l
tbr_l	boarding time on link l
td_l	dwelling time on link l
tw_{lj}	waiting time on diversion link l for hyperpath j
t_l	total travel time for link l
tr_l	running time for link l
tw_l	average delay (at intersection) for link l
$\boldsymbol{U}^i$	vector of perceived utilities for user i
U^i_j	perceived utility of alternative j for user i
v_l	average speed on link l
$\boldsymbol{V}^i$	vector of systematic utilities for user i
V^i_j	systematic utility of alternative j for user i
$\boldsymbol{V}_{od}$	vector of the systematic utilities V_k of the od pair
$\boldsymbol{V}_{od,i}$	vector of the systematic utilities of the od pair for the class i
$\boldsymbol{V}_{od,m}$	vector of systematic utilities for paths related to the od pair and the mode m
x_j	cost of hyperpath j
$\boldsymbol{x}$	vector of total hyperpath costs
x_j^{ADD}	additive cost of hyperpath j
$\boldsymbol{x}^{ADD}$	vector of hyperpath additive costs
x_j^{NA}	non additive cost of hyperpath j
$\boldsymbol{x}^{NA}$	vector of hyperpath non-additive costs
$\boldsymbol{x}_{od}$	vector of total hyperpath costs for the users of the od pair
$\boldsymbol{x}_{od}^{NA}$	vector of hyperpath non-additive costs for the users of the od pair
$\boldsymbol{x}_{od}^{ADD}$	vector of hyperpath additive costs for the users of the od pair
y_j	flow on hyperpath j
$\boldsymbol{y}$	vector of hyperpath flows
$\boldsymbol{y}_{od}$	vector of hyperpath flows for the od pair
z_{nk}	path performance variable n for path k
$z^{ADD}{}_{nk}$	additive path performance variable n for path k
$z^{NA}{}_{nk}$	non-additive path performance variable n for path k
Z_{od}	minimum (shortest path) cost between pairs o and d
$\boldsymbol{Z}$	vector of the minimum path costs between all the od pairs
β_k	coefficient of the k^{th} attribute in systematic utility
$\boldsymbol{\beta}$	vector of coefficients in systematic utility
δ_{lk}	link-path incidence element
$\boldsymbol{\Delta}$	link-path incidence matrix
$\boldsymbol{\Delta}_{od}$	link-path incidence matrix for od pair
$\boldsymbol{\Delta}_{od,i}$	link-path incidence matrix for the od pair and the class i
$\boldsymbol{\Delta}_{od,m}$	link-path incidence matrix for the od pair and the mode m
δ_r	hierarchical logit coefficient associated to intermediate node r
η_{lj}	diversion probability on link l within hyperpath j

η^i_j	component random residuals of alternative j for user i
θ_r	generic coefficient of the Gumbel random variable
$\boldsymbol{\theta}$	vector of coefficients of Gumbel random variables
λ_{lj}	probability of crossing link l within hyperpath j
$\lambda_{od,lj}$	probability of traversing link l within the hyperpath j for the users of the od pair
$\boldsymbol{\Lambda}$	matrix of the probabilities of traversing the links within each hyperpath
$\boldsymbol{\Lambda}_{od}$	matrix of the traversing probabilities $\lambda_{od,lj}$ of each link l within each hyperpath j for the users of the od pair
Σ_x	variance-covariance (or dispersion) matrix of the vector of random variable x
$\varphi_l\ (\varphi_{mn})$	frequency of transit line accessed through link l (m,n)
Φ_m	cumulated frequency at diversion node m
Φ_{mj}	cumulated frequency of the lines accessible at diversion mode m in the hyperpath j
ω_{kj}	probability of path k within the hyperpath j
$\omega_{od,kj}$	probability of choosing path k within the hyperpath j for a user of the od pair
$\boldsymbol{\Omega}$	path-hyperpath probability matrix
$\boldsymbol{\Omega}_{od}$	matrix of the path choice probabilities $\omega_{od,kj}$ within the hyperpaths for the od pair

Applied Optimization

18. O. Maimon, E. Khmelnitsky and K. Kogan: *Optimal Flow Control in Manufacturing*. Production Planning and Scheduling. 1998 ISBN 0-7923-5106-1
19. C. Zopounidis and P.M. Pardalos (eds.): *Managing in Uncertainty: Theory and Practice*. 1998 ISBN 0-7923-5110-X
20. A.S. Belenky: *Operations Research in Transportation Systems:* Ideas and Schemes of Optimization Methods for Strategic Planning and Operations Management. 1998 ISBN 0-7923-5157-6
21. J. Gil-Aluja: *Investment in Uncertainty*. 1999 ISBN 0-7923-5296-3
22. M. Fukushima and L. Qi (eds.): *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smooting Methods*. 1999 ISBN 0-7923-5320-X
23. M. Patriksson: *Nonlinear Programming and Variational Inequality Problems*. A Unified Approach. 1999 ISBN 0-7923-5455-9
24. R. De Leone, A. Murli, P.M. Pardalos and G. Toraldo (eds.): *High Performance Algorithms and Software in Nonlinear Optimization*. 1999 ISBN 0-7923-5483-4
25. A. Schöbel: *Locating Lines and Hyperplanes*. Theory and Algorithms. 1999 ISBN 0-7923-5559-8
26. R.B. Statnikov: *Multicriteria Design*. Optimization and Identification. 1999 ISBN 0-7923-5560-1
27. V. Tsurkov and A. Mironov: *Minimax under Transportation Constrains*. 1999 ISBN 0-7923-5609-8
28. V.I. Ivanov: *Model Development and Optimization*. 1999 ISBN 0-7923-5610-1
29. F.A. Lootsma: *Multi-Criteria Decision Analysis via Ratio and Difference Judgement*. 1999 ISBN 0-7923-5669-1
30. A. Eberhard, R. Hill, D. Ralph and B.M. Glover (eds.): *Progress in Optimization*. Contributions from Australasia. 1999 ISBN 0-7923-5733-7
31. T. Hürlimann: *Mathematical Modeling and Optimization*. An Essay for the Design of Computer-Based Modeling Tools. 1999 ISBN 0-7923-5927-5
32. J. Gil-Aluja: *Elements for a Theory of Decision in Uncertainty*. 1999 ISBN 0-7923-5987-9
33. H. Frenk, K. Roos, T. Terlaky and S. Zhang (eds.): *High Performance Optimization*. 1999 ISBN 0-7923-6013-3
34. N. Hritonenko and Y. Yatsenko: *Mathematical Modeling in Economics, Ecology and the Environment*. 1999 ISBN 0-7923-6015-X
35. J. Virant: *Design Considerations of Time in Fuzzy Systems*. 2000 ISBN 0-7923-6100-8

Applied Optimization

36. G. Di Pillo and F. Giannessi (eds.): *Nonlinear Optimization and Related Topics.* 2000 ISBN 0-7923-6109-1
37. V. Tsurkov: *Hierarchical Optimization and Mathematical Physics.* 2000 ISBN 0-7923-6175-X
38. C. Zopounidis and M. Doumpos: *Intelligent Decision Aiding Systems Based on Multiple Criteria for Financial Engineering.* 2000 ISBN 0-7923-6273-X
39. X. Yang, A.I. Mees, M. Fisher and L.Jennings (eds.): *Progress in Optimization.* Contributions from Australasia. 2000 ISBN 0-7923-6175-X
40. D. Butnariu and A.N. Iusem: *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization.* 2000 ISBN 0-7923-6287-X
41. J. Mockus: *A Set of Examples of Global and Discrete Optimization.* Applications of Bayesian Heuristic Approach. 2000 ISBN 0-7923-6359-0
42. H. Neunzert and A.H. Siddiqi: *Topics in Industrial Mathematics.* Case Studies and Related Mathematical Methods. 2000 ISBN 0-7923-6417-1
43. K. Kogan and E. Khmelnitsky: *Scheduling: Control-Based Theory and Polynomial-Time Algorithms.* 2000 ISBN 0-7923-6486-4
44. E. Triantaphyllou: *Multi-Criteria Decision Making Methods.* A Comparative Study. 2000 ISBN 0-7923-6607-7
45. S.H. Zanakis, G. Doukidis and C. Zopounidis (eds.): *Decision Making: Recent Developments and Worldwide Applications.* 2000 ISBN 0-7923-6621-2
46. G.E. Stavroulakis: *Inverse and Crack Identification Problems in Engineering Mechanics.* 2000 ISBN 0-7923-6690-5
47. A. Rubinov and B. Glover (eds.): *Optimization and Related Topics.* 2001 ISBN 0-7923-6732-4
48. M. Pursula and J. Niittymäki (eds.): *Mathematical Methods on Optimization in Transportation Systems.* 2000 ISBN 0-7923-6774-X
49. E. Cascetta: *Transportation Systems Engineering: Theory and Methods.* 2001 ISBN 0-7923-6792-8

KLUWER ACADEMIC PUBLISHERS – DORDRECHT / BOSTON / LONDON