ECONOMICS OF AGGLOMERATION

CITIES, INDUSTRIAL LOCATION, AND REGIONAL GROWTH

MASAHISA FUJITA
JACQUES-FRANÇOIS THISSE

CAMBRIDGE www.cambridge.org/9780521801386

This page intentionally left blank

Economics of Agglomeration

This book provides the first unifying treatment of the range of economic reasons for the clustering of firms and households. Its goal is to explain further the trade-off between various forms of increasing returns and different types of mobility costs. It should be noted that the concept of economic agglomeration refers to very distinct real-world situations. The main focus of the treatment is on cities, but *Economics of Agglomeration* also explores the formation of commercial districts within cities, industrial clusters at the regional level, and the existence of imbalance between regions. The book is rooted within the realm of modern economics and borrows concepts and ideas from geography and regional science, which makes it accessible to a broad audience comprising economists, geographers, regional planners, and other scientists. *Economics of Agglomeration* may be used in coursework for graduate students and talented upper-level undergraduates.

Masahisa Fujita has been a major contributor to spatial economic theory during his 20-year tenure at the University of Pennsylvania and more recently at Kyoto University since 1995, where he has served as Director of the Institute of Economic Research. His scholarship ranges over the fields related to regional science, location theory, economic geography, urban economics, and international trade. Professor Fujita is the author or coauthor of three books: *Spatial Development Planning* (1978), *Urban Economic Theory* (Cambridge University Press, 1989), which remains to this day the most authoritative graduate textbook on urban economics, and *The Spatial Economy* (1999, coauthored with Paul Krugman and A.J. Venables), which defines the field of New Economic Geography. Professor Fujita is the recipient of the 1983 Tord Palander Prize, the 1998 Walter Isard Award in regional sciences, and the 2000 Nikkei Economic Book Prize. In 1993 he was honored with the lifetime Visiting Professorship in the College of Economics at Nankai University, China. He serves on the board of 10 international journals on regional science, economics, and economic geography and since 1999 has been the President of the Applied Research Science Conference in Japan.

Jacques-François Thisse is Professor of Economics at the Université catholique de Louvain, President of the Center for Operations Research and Econometrics (CORE), and Professor at the Ecole Nationale des Ponts et Chaussées. A Fellow of the Econometric Society, he is an internationally eminent applied theorist in industrial organization, economic geography, and public economics. He is the author of more than 150 articles in academic journals, including *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Review of Economic Studies*, *Journal of Economic Theory*, *Journal of Urban Economics*, *Journal of Public Economics*, *Operations Research*, *Management Science*, *Journal of Mathematical Psychology*, and *Geographical Analysis*. He has been Professor at the Sorbonne and head of CERAS (Paris) as well as a visiting professor at McMaster University, Canada, University of Pennsylvania, INSEAD, and Virginia Polytechnic Institute. He coauthored *Discrete Choice Theory of Product Differentiation* (1992) and coedited *Economics of Cities* (1999) published by Cambridge University Press. He was awarded the Palander Prize in Regional Science in 1986 and has served on the editorial board of more than 10 different academic journals.

Economics of Agglomeration
Cities, Industrial Location, and Regional Growth

MASAHISA FUJITA
Kyoto University

JACQUES-FRANÇOIS THISSE
Université catholique de Louvain

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521801386

First published in print format 2002

ISBN-13 978-0-521-80138-6 Hardback

ISBN-13 978-0-521-80524-7 Paperback

To Yuko
To Ginou

Primitive though it may be, every stable society feels the need of providing its members with centers of assembly, or meeting places. Observance of religious rites, maintenance of markets, and political and judicial gatherings necessarily bring about the designation of localities intended for the assembly of those who wish to or who must participate therein.

Henri Pirenne, *Medieval Cities* (1925)

Contents

Acknowledgments

This book is the outcome of joint research that was started several years ago. Our interest in urban economics and location theory stems from early work developed independently with friends to whom we have an intellectual debt. We should like to thank them – in particular Jean Gabszewicz and Tony Smith.

During the preparation of the manuscript, we were fortunate to have had the opportunity to discuss the subject matter with many colleagues, friends, and students; some of them also read earlier drafts of several chapters with care and made detailed and useful suggestions.

We would like to mention especially Marcus Berliant, Luisito Bertinelli, Raouf Boucekkine, Jacques Drèze, Gilles Duranton, Philippe Martin, Tomoya Mori, Gianmarco Ottaviano, Dominique Peeters, Pierre Picard, David Pines, Hiroshi Sano, Takatoshi Tabuchi, Tanguy van Ypersele, Luigi Ventura, Kazuhiro Yamamoto, and Dao-Zhi Zeng. We also owe a special debt to Diego Puga who performed superbly as an "anonymous" referee. Thanks are also due to two anonymous referees for very helpful comments.

The final draft of the manuscript was prepared at CORE by Sylvie Mauroy, and the figures were composed at Kyoto University by Yasuko Kamiyoshi and Tomoko Nomura. We want to express our gratitude for their patience and competence.

Finally, we acknowledge the financial support of the Japanese Ministry of Education and Science (Grant-in-Aid for Science Research 09CE2002 and 13851002), the Fonds National de la Recherche Scientifique (Belgium) for a grant to the second author (Convention 2.4529.98F), and the Center for Operations Research and Econometrics (CORE) at the Université catholique de Louvain. We also thank CERAS at the Ecole Nationale des Ponts et Chaussées (Paris) and the Faculty of Commerce and Business Administration of the University of British Columbia for their friendly hospitality.

1

Agglomeration and Economic Theory

1.1 INTRODUCTION

Just as matter in the solar system is concentrated in a small number of bodies (the planets and their satellites), economic life is concentrated in a fairly limited number of human settlements (cities and clusters). Furthermore, paralleling large and small planets, there are large and small settlements with very different combinations of firms and households. This book is a study of the reasons for the existence of a large variety of economic agglomerations. Even though economic activities are, to some extent, spatially concentrated because of natural features (think of rivers and harbors), our goal is to focus on economic mechanisms yielding agglomeration by relying on the trade-off between various forms of increasing returns and different types of mobility costs.

One should keep in mind that the concept of economic agglomeration refers to very distinct real-world situations.[1] At one extreme lies the core–periphery structure corresponding to North–South dualism. For example, Hall and Jones (1999) observed that high-income nations are clustered in small industrial cores in the Northern Hemisphere and that productivity per capita steadily declines with distance from these cores.

As noted by many historians and development theorists, economic growth tends to be localized. This is especially well illustrated by the rapid growth of East Asia during the last few decades. We view East Asia here as comprising Japan and nine other countries, that is, Republic of South Korea, Taiwan, Hong Kong, Singapore, Philippines, Thailand, Malaysia, Indonesia, and China. In 1990, the total population of East Asia was about 1.6 billion. With only 3.5% of the total area and 7.9% of the total population, Japan accounted for 72% of the gross domestic product (GDP) and 67% of the manufacturing GDP of East Asia. In Japan itself, the economy is very much dominated by its core regions formed by the five prefectures containing the three major metropolitan areas of Japan: Tokyo and Kanagawa prefectures, Aichi prefecture (containing the Nagoya metropolitan area), and Osaka and Hyogo prefectures. These regions account for only 5.2% of the area of Japan but for 33% of its population, 40% of its GDP, and 31% of its manufacturing employment. Hence, for the whole of East Asia, the Japanese core regions with a mere 0.18% of the total area accounted for 29% of East Asia's GDP.

Strong regional disparities within the same country imply the existence of agglomerations at another spatial scale. For example, in Korea, the capital region (Seoul and Kyungki Province), which has an area corresponding to 11.8% of the country and includes 45.3% of the population, produces 46.2% of the GDP. In France, the contrast is even greater: the Île-de-France (the metropolitan area of Paris), which accounts for 2.2% of the area of the country and 18.9% of its population, produces 30% of its GDP. Inside the Île-de-France, only 12% of the available land is used for housing, plants, and roads, the remaining land being devoted to agriculture, forestry, or natural activities.

Regional agglomeration is also reflected in large varieties of cities, as shown by the stability of the urban hierarchy within most countries (J. Eaton and Eckstein 1997; Dobkins and Ioannides 2000). Cities themselves may be specialized in a very small number of industries, as are many medium-size American cities (Henderson 1997a). However, large metropolises like New York or Tokyo are highly diversified in that they nest many industries that are not related through direct linkages (Chinitz 1961; Fujita and Tabuchi 1997). Industrial districts involving firms with strong technological, or informational linkages, or both (e.g., the Silicon Valley or Italian districts engaged in more traditional activities) as well as factory towns (e.g., Toyota City or IBM in Armonk, New York) manifest various types of local specialization. Therefore, it appears that highly diverse size and activity arrangements exist at the regional and urban levels.

At a very detailed extreme of the spectrum, agglomeration arises under the form of large commercial districts set up in the inner city itself (think of Soho in London, Montparnasse in Paris, or Ginza in Tokyo). At the lowest level, restaurants, movie theaters, or shops selling similar products are clustered within the same neighborhood, not to say on the same street, or the clustering may take the form of a large shopping mall. Understanding such phenomena is critical for the design of effective urban policies.

The economic reasons that stand behind such strong geographical concentrations of consumption and production are precisely what we aim to investigate in this book. To achieve this objective, we will appeal to the concepts and tools of modern microeconomics. Because clusters appear at different geographical scales and involve various degrees of sectoral details, it would be futile to look for *the* model explaining different types of economic agglomerations (Papageorgiou 1983). This should not come as a surprise, for geographers have long known that geographical scale matters.[2] What is true at a certain spatial scale is not necessarily true at another (the "ecological fallacy"). For example, whether Los Angeles or Chicago may be considered as a megacenter or as a collection of several large subcenters depends very much on the scale of observation. Likewise, during the 1980s the income differentials have decreased across country members of the European Union but not across regions within countries. The reason for such differences probably lies in the nature and balance of the system of forces at work at a given level of analysis. Or, in the words of Anas, Arnott, and Small (1998, 1440):

> It may be that the patterns that occur at different distance scales are influenced by different types of agglomeration economies, each based on interaction mechanisms with particular requirements for spatial proximity.

Yet, as will be seen, a few general principles seem to govern the formation of distinct agglomerations even though the content and intensity of the forces at work may vary with place and time.[3]

1.2 CITIES: PAST AND FUTURE

Casual observation reveals the extreme variation in the intensity of human settlements and land use – a fact that has culminated in the existence of *cities* in which population densities are very high.[4]

From a historical perspective, cities emerged in several parts of the world about 7,000 years ago as the consequence of the rise in agricultural surplus. The mere existence of cities may be viewed as a universal phenomenon whose importance slowly but steadily increased during the centuries preceding the sudden urban growth that appeared during the nineteenth century in a small corner of Europe (Bairoch 1985, chaps. 15–17). Technological development was necessary to generate the agricultural surplus without which cities would have been inconceivable at the time, as they would be today.

In addition to technological innovations, a fundamental change in social structure was also necessary: the division of labor into specialized activities. In this respect, there seems to be a large agreement among economists, geographers, and historians to consider "increasing returns" as the most critical factor in the emergence of cities. For example, J. Marshall (1989, 25) has suggested that

> quite apart from considerations related to defense, to royal whim, or to the supposed sacred importance of certain sites, the formation of towns made good economic sense in promoting a level of efficiency in commerce, manufacturing, and administration that would have been impossible to achieve with a completely dispersed population.

Although the sources are dispersed, not always trustworthy, and hardly comparable, data clearly converge to show the existence of an urban revolution. In Europe, the proportion of the population living in cities increased very slowly from 10% in 1300 to 12% in 1800 (Bairoch 1985). It was approximately 20% in 1850, 38% in 1900, 52% in 1950, and is now close to 75%, thus showing an explosive growth in the urban population (Bairoch 1985; United Nations 1994). In the United States, the rate of urbanization increased from 5% in 1800 to more than 60% in 1950 and is now nearly 77%. In Japan, the rate of urbanization was about 15% in 1800 (Bairoch 1985), 50% in 1950, and is now about 78% (United Nations 1994). The proportion of the urban population in the world increased from 30% in 1950 to 45% in 1995 and will exceed 50% in 2005 (United Nations 1994). The world's urban population increases each year by the equivalent of 40 million (i.e., the population of Spain).

Furthermore, concentration in very big cities keeps rising. In 1950, only two cities had populations greater than 10 million: New York and Greater London. In 1995, fifteen cities belonged to this category. The largest one, Tokyo, with more than 26 million, exceeds the second one, New York, by 10 million. In 2025, 26 megacities will exceed 10 million in population (United Nations 1994).

Economists and geographers must explain why firms and households concentrate in large metropolitan areas even though empirical evidence suggests that the cost of living in such areas is typically higher than in smaller urban areas (Richardson 1987). As Lucas (1988, 39) neatly put it, "What can people be paying Manhattan or downtown Chicago rents for, if not for being near other people?" But Lucas did not explain why people want, or need, to be near other people. Likewise, economists and geographers must explain the formation of small and specialized clusters of firms and workers not necessarily located within major cities – such as many of the Italian industrial districts (Pyke, Becattini, and Sengenberger 1990, chap. 3) – and that appear to be very efficient in terms of productivity.

The increasing availability of high-speed transportation infrastructure and the fast-growing development of new informational technologies might suggest that our economies are entering an age that will culminate in the "death of distance." If so, locational difference would gradually fade because agglomeration forces would be vanishing. In other words, cities would become a thing of the past. We will see in this book that things are not that simple because the opposite trend may just as well arise. Indeed, one of the general principles to be derived from our analysis is that the relationship between the decrease in transport costs and the degree of agglomeration of economic activities is not that expected by many analysts: *Agglomeration happens provided that transport costs are below some critical threshold*,[5] although further decreases may yield dispersion of some activities owing to factor price differentials. In addition, technological progress brings about new types of innovative activities that benefit most from being agglomerated and, therefore, tend to arise in developed areas. Consequently, the wealth or poverty of nations seems to be more and more related to the development of prosperous and competitive clusters of specific industries as well as to the existence of large and diversified metropolitan areas (Glaeser 1998; Porter 1998, chaps. 6 and 7; Thisse and van Ypersele 1999).

The recent attitude taken by several institutional bodies and medias seems to support this view. For example, in its recent *World Development Report*, the World Bank (2000) stressed the importance of economic agglomerations and cities for boosting growth and escaping from the poverty trap. Another example of this increasing awareness of the relevance of cities in modern economies can be found in *The Economist* (1995, 18):

> The liberalization of world trade and the influence of regional trading groups such as NAFTA and the EU will not only reduce the powers of national governments, but also increase those of cities. This is because an open trading system will have the effect of making national economies converge, thus evening out the competitive advantage of countries, while leaving those of cities largely untouched. So in the future, the arenas in which companies will compete may be cities rather than countries.

In this book, we intend to address the main causes for the formation of the various types of economic agglomerations described above. As discussed in the next two sections, this includes increasing returns to scale, externalities, and imperfectly competitive markets with general and strategic interdependencies. From this list, it should be clear that the economics of agglomeration is fraught with most of the difficulties encountered in economic theory.

Moreover, as will be seen in various chapters of this book, models of agglomeration involve both *complementarity* and *substitution* effects. For a long time, economists had problems handling complementarity effects, which can hardly be taken in account in the general competitive framework. This observation will lead us, in Section 1.4, to survey the rather complex history of the relationship between space and economic theory. Although space has not been ignored by some prominent economists, it has seldom been mentioned in economics texts. Thus, it is interesting to determine why this important ingredient of social life has been put aside for so long.

1.3 WHY DO WE OBSERVE AGGLOMERATIONS?

Intuitively, it should be clear that the spatial configuration of economic activities is the outcome of a process involving two opposing types of forces, that is, *agglomeration* (or centripetal) forces and *dispersion* (or centrifugal) forces. The observed spatial configuration of economic activities is then the result of a complicated balance of forces that push and pull consumers and firms. This view agrees with very early work in economic geography. For example, in his *Principes de géographie humaine* published posthumously in 1921, the famous French geographer Vidal de la Blache argued that all societies, rudimentary or developed, face the same dilemma: Individuals must get together to benefit from the advantages of the division of labor, but various difficulties restrict the gathering of many individuals.

1.3.1 Agglomeration and Increasing Returns

One would expect trade theory to be the branch of economics that has paid most attention to the spatial dimension. The reason is that changes in the conditions under which commodities are shipped, as well as changes in the mobility of factors, affect the location of industry, the geography of demand and, eventually, the pattern of trade. The opposite has been true, for neoclassical trade theory has treated each country as dimensionless and has given little attention to the impact of trade costs. Yet, some predominant contributors in the field have long argued that location and trade are closely related topics. For example, Ohlin (1933; 1968, 97) has challenged the common wisdom that considers international trade theory as separate from location theory:[6]

> International trade theory cannot be understood except in relation to and as part of the general location theory, to which the lack of mobility of goods and factors has equal relevance.

Natural resources, and more generally production factors, are not uniformly distributed across locations, and it is on this unevenness that most of trade theory has been built.[7] The standard model of trade considers a setting formed by two countries producing two goods by means of two factors (labor and capital) under identical technologies subject to constant returns to scale and strictly diminishing marginal products. When factors are spatially immobile and goods can be costlessly moved from one country to the other, this model predicts the equalization of factor prices when the ratios of factor endowments are not too different.

Similarly, regional economics has long been dominated by the dual version of the neoclassical trade model. It is assumed that a single good is produced and that (at least) one production factor can *freely* move between regions. According to this model, capital flows from regions where it is abundant to regions where it is scarce until capital rents are the same across regions, or regional wage differences push and pull workers until the equalization of wages between regions is reached. Because the production function is linear homogeneous and has strictly diminishing marginal product in each factor, the marginal productivity of the mobile factor depends only on the capital–labor ratio. This implies that the mobile factor moves from regions with low returns toward regions with high returns up to the point at which the capital–labor ratio is equalized across all regions. In other words, the perfect mobility of one factor would be sufficient to guarantee the equalization of wages and capital rents in the interregional marketplace.[8]

Thus, it would seem that either costless trade or the perfect mobility of one factor would be sufficient to guarantee the convergence of labor income across various places.[9] Ignoring unevenness in the spatial distribution of natural resources, Mills (1972a, 4) very suggestively described this strange "world without cities" that would characterize an economy operating under constant returns and perfect competition as follows:

> Each acre of land would contain the same number of people and the same mix of productive activities. The crucial point in establishing this result is that constant returns

permit each productive activity to be carried on at an arbitrary level without loss of efficiency. Furthermore, all land is equally productive and equilibrium requires that the value of the marginal product, and hence its rent, be the same everywhere. Therefore, in equilibrium, all the inputs and outputs necessary directly and indirectly to meet the demands of consumers can be located in a small area near where consumers live. In that way, each small area can be autarkic and transportation of people and goods can be avoided.

Such an economic space is the quintessence of self-sufficiency. This suggests, therefore, that the constant returns–perfect competition paradigm is unable to cope with the emergence and growth of large economic agglomerations (Krugman 1995, chap. 1).

Increasing returns in production activities are needed if we want to explain economic agglomerations without appealing to the attributes of physical geography. In particular, the trade-off between increasing returns in production and transportation costs is central to the understanding of the geography of economic activities. Although it has been rediscovered many times (including in recent periods), this idea has been at the heart of the work developed by early location theorists. For example, Lösch ([1940] 1954) stated that:

We shall consider market areas that are not the result of any kind of natural or political inequalities but arise through the interplay of purely economic forces, some working toward concentration, and others toward dispersion. In the first group are the advantages of specialization and of large-scale production; in the second, those of shipping costs and of diversified production (p. 105 of the English translation).

It is only during the 1990s that some trade theorists became aware that "they were doing geography without knowing it" and have turned their attention to spatial issues. Since then, it is fair to say that they have contributed significantly in promoting geographical economics through the use of models involving both monopolistic competition and increasing returns (Krugman 1991a,b; Venables 1996; Helpman 1998).[10]

1.3.2 Agglomeration and Externalities

According to A. Marshall ([1890], 1920, chap. X), externalities are crucial in the formation of economic agglomerations and generate something like a lock-in effect:

When an industry has thus chosen a location for itself, it is likely to stay there long: so great are the advantages which people following the same skilled trade get from near neighbourhood to one another. The mysteries of the trade become no mysteries; but are as it were in the air, and children learn many of them unconsciously. Good work is rightly appreciated, inventions and improvements in machinery, in processes and the general organization of the business have their merits promptly discussed: if one man starts a new idea, it is taken up by others and combined with suggestions of their own; and thus it becomes the source of further new ideas (p. 225).

For this author, relevant externalities for the formation of clusters involve the following:

1. mass production (the internal economies that are identical to scale economies at the firm's level);
2. availability of specialized input services;
3. formation of a highly specialized labor force and the production of new ideas, both based on the accumulation of human capital and face-to-face communications; and
4. the existence of modern infrastructure.[11]

Despite its vagueness, the concept of Marshallian externalities has been much used in the economics and regional science literature devoted to the location of economic activities because it captures the idea that an agglomeration is the outcome of a "snowball effect" in which a growing number of agents want to congregate to benefit from a larger diversity of activities and a higher specialization.[12] Such cumulative processes are now associated with the interplay of pecuniary externalities in models combining increasing returns and monopolistic competition (Matsuyama 1995).[13]

In fact, the concept of externality has been used to describe a great variety of situations. Following Scitovsky (1954), it is now customary to consider two categories: "technological externalities" (also called spillovers) and "pecuniary externalities." The former deals with the effects of nonmarket interactions that are realized through processes directly affecting the utility of an individual or the production function of a firm. In contrast, pecuniary externalities are by-products of market interactions: They affect firms or consumers and workers only insofar as they are involved in exchanges mediated by the price mechanism. Pecuniary externalities are relevant when markets are imperfectly competitive, for when an agent's decision affects prices, it also affects the well-being of others.

According to Anas et al. (1998), cities would be replete with technological externalities. The same would hold in local production systems (Pyke et al. 1990, chap. 4). In fact, much of the competitiveness of individuals and firms is due to their creativity, and thus economic life is creative in the same way as are the arts and sciences. Of particular interest for creativity are "communication externalities." This idea accords with the view of Lucas (1988, 38) when he writes that "New York City's garment district, financial district, diamond district, advertising district and many more are as much intellectual centers as is Columbia or New York University." Thus, to explain geographical clusters of somewhat limited spatial dimension such as cities and highly specialized industrial and scientific districts, it seems reasonable to appeal to technological externalities, which, in terms of modeling, have the additional advantage of being compatible with the competitive paradigm.

The advantages of proximity for production have their counterpart on the consumption side. For example, the propensity to interact with others is a fundamental human attribute, as is the tendency to derive pleasure in discussing and exchanging ideas with others. Distance is an impediment to such interactions, and thus cities are the ideal institution for the development of social contacts. Along the same line, Akerlof (1997) argued that the inner city is often the substratum for the development of social norms such as conformity and status seeking that govern the behavior of groups of agents.

On the other hand, when we consider a large geographical area, it seems reasonable to think that direct physical contact provides a weak explanation of interregional agglomerations such as the "Manufacturing Belt" in the United States and the "Blue Banana" in Europe (an area that stretches from London to northern Italy and goes through part of western Germany and the Benelux countries). This is the realm of pecuniary externalities that arise from imperfect competition in the presence of market-mediated linkages between firms and consumers and workers. Such externalities lie at the heart of models of monopolistic competition recently developed to explain the agglomeration of economic activities; they also have one major intellectual advantage.

To a large extent, technological externalities are often black boxes that aim at capturing the crucial role of complex nonmarket institutions whose role and importance are strongly stressed by geographers and spatial analysts (see, e.g., Pyke et al. 1990; Saxenian 1994). By contrast, because pecuniary externalities focus on economic interactions mediated by the market, their origin is clearer. In particular, their impact can be traced back to the values of fundamental microeconomic parameters such as the intensity of returns to scale, the strength of firms' market power, the level of barriers to goods, and factor mobility.

Whatever externalities are at work, prices do not fully reflect the social values of goods and services, and thus market outcomes are likely to be inefficient. The dominant feeling in the economics profession is that most cities and agglomerations are just too big. The prevalence of big and gloomy slums in Third World megalopolises gives the impression that the laissez-faire policy has led to an excessive concentration of human beings in excessively large agglomerations all over the world. Likewise, most regional policy debates in industrialized countries implicitly assume that there is too much spatial concentration. In this respect, Hotelling (1929, 57) stated more than 70 years ago what probably remains the conventional wisdom of economists regarding cities and the spatial organization of economic activities: "Our cites become uneconomically large and the business districts within them are too concentrated." We will see in this book that things are not that simple. Urban externalities are not necessarily negative, and increasing returns might be a strong force in favor of geographical concentration. Hence, it seems fair to say that there is no presumption regarding the direction in which governments should move in their regional and urban policies.[14]

1.3.3 Thünen and Agglomerations

At this stage, it is worth noting that the economics profession has ignored the previous availability in Thünen's work of most of the factors explaining economic agglomerations.[15] When asking whether industrial firms are better off located in major cities (especially in the capital), Thünen ([1826] 1966) started by describing the main centrifugal forces at work:

> 1. Raw materials are more expensive than in the country towns on account of the higher cost of transport. 2. Manufactured articles incur the cost of haulage to the provincial towns when they are distributed to the rural consumers. 3. All necessities, especially firewood, are much more expensive in the large town. So is rent for flats and houses, for two reasons (1) construction costs are higher because raw materials have to be brought from a distance and are consequently more expensive, and (2) sites that may be bought for a few thalers in a small town are very dear. Since food, as well as fuel and housing, cost so much more in the large town, the wage expressed in money, must be much higher than in the small one. This adds appreciably to production costs (pp. 286–7 of the English translation).

This list is surprisingly comprehensive. In particular, the impact of high land rents and high food prices on monetary wages in large cities is explicitly spelled out (see Chapter 6).

Thünen then turned to the centripetal forces that, according to him, stand behind industrial agglomerations.

> 1. Only in large-scale industrial plants is it profitable to install labour-saving machinery and equipment, which economise on manual labour and make for cheaper and more efficient production. 2. The scale of an industrial plant depends on the demand for its products. . . . 4. For all these reasons, large scale plants are viable only in the capital in many branches of industry. But the division of labour (and Adam Smith has shown the immense influence this has on the size of the labour product and on economies of production) is closely connected with the scale of an industrial plant. This explains why, quite regardless of economies of machine-production, the labour product per head is far higher in large than in small factories. . . . 7. Since it takes machines to produce machines, and these are themselves the product of many different factories and workshops, machinery is produced efficiently only in a place where factories and workshops are close enough together to help each other work in unison, i.e. in large towns. . . . Economic theory has failed to adequately appreciate this factor. Yet it is this which explains why factories are generally found communally, why, even when in all other respects conditions appear suitable, those set up by themselves, in isolated places, so often come to grief. Technical innovations are continually increasing the complexity of machinery; and the more complicated the machines, the more the factor of association will enter into operation (pp. 287–90 of the English translation).

Observe that the combination of Thünen's agglomeration factors 1, 2, and 4 almost coincides with Krugman's "basic story" for the emergence of a core–periphery structure (see Chapter 9). Furthermore, if we combine these factors with the last one (7), which is about interindustry linkages and technological spillovers, we get another fundamental explanation for the emergence of industrial agglomerations (see Chapters 7 and 9).

Even though Thünen's work took place at the very beginning of the Industrial Revolution in Germany, it would be hard to imagine a more explicit description of the forces shaping the industrial landscape.

1.4 ON THE RELATIONSHIP BETWEEN SPACE AND ECONOMICS

It is rare to find an economics text in which space is studied as an important subject – if it is even mentioned. As argued by Krugman (1995, chap. 1), this is probably because economists lacked a model embracing both increasing returns and imperfect competition, the two basic ingredients of the formation of the economic landscape, as shown by the pioneering work of Hotelling (1929), Lösch (1940), Isard (1956), Koopmans (1957), and Greenhut (1963).[16]

Certainly many eminent economists have turned their attention to the subject at least in passing, and Samuelson (1983) places the subject's founder, Thünen, in the pantheon of great economists. Thünen ([1826] 1966) sought to explain the pattern of agricultural activities surrounding a typical city in preindustrial Germany, and we will see that his theory has proven to be very useful in studying land use when economic activities are perfectly divisible. In fact, the principles underlying his model are so general that Thünen can be considered the founder of marginalism (Samuelson 1983; Nerlove and Sadka 1991). Ekelund and Hébert (1999, 246) go one step further when they claim that "With uncommon brilliance and deftness Thünen virtually invented the modern economic 'model,' which integrates logical deduction with factual experiment." In addition, the import of Thünen's analysis for the development of geographical economics is twofold in that space is considered as both an economic good and as the substratum for economic activities, thus making his work more relevant and general than several later contributions.

Despite his monumental contribution to economic thought, Thünen's ideas languished for more than a century without attracting widespread attention. Why was this so? According to Ekelund and Hébert (1999, 245), the reason lies in the work and influence of Ricardo:

> The economics of David Ricardo constituted a negative watershed in the history of spatial theory. By reducing situational differences to differences in the fertility of land, Ricardo effectively eliminated spatial considerations from his analytical system. Moreover, he made transportation costs indistinguishable from other costs, and in international trade theory where spatial considerations had previously dominated, he substituted comparative costs as the crucial factor. The practical effect of Ricardo's method and of his analytical innovations was to dislodge space from mainstream economic theory, so that for a long period thereafter it came to be treated, if at all, outside the mainstream deductive models of British classical economics.

Aside from such an unfortunate historical whim, Thünen's theory left a crucial issue unexplored: Why is there a city in Thünen's isolated state? Although such a center may emerge under constant returns when space is heterogeneous (Beckmann and Puu 1985), a city is more likely to arise when increasing returns are at work in the design of trading places or in the production of some goods. In other words, one must appeal to "something" that is not in the Thünian model to understand what is going on.

There is an interesting analogy between the Thünen's model and Solow's (1956) growth model. Both assume constant returns to scale and perfect competition. As in Thünen's, in which the city cannot be explained within the model, the main reason for growth, that is, technological progress, cannot be explained within the model of exogenous growth. This difficulty is well summarized by Romer (1992, 85–6) in the following paragraph:

> The paradox . . . was that the competitive theory that generated the evidence was inconsistent with any explanation of how technological change could arise as the result of the self-interested actions of individual economic actors. By definition, all of national output had to be paid as returns to capital and labor; none remained as possible compensation for technological innovations. . . . The assumption of convexity and perfect competition placed the accumulation of new technologies at the center of the growth process and simultaneously denied the possibility that economic analysis could have anything to say about this process.

Stated differently, explaining city formation in Thünian models is similar to explaining technological progress in the neoclassical growth model.

Despite this limitation, the Thünian model has proven its relevance lately for the development of spatial economics. Following the suggestion made by Isard (1956, chap. 8), Alonso (1964) succeeded in extending Thünen's central concept of bid rent curves to an urban context in which a marketplace is replaced by an employment center (the Central Business District). Since that time urban economics has advanced rapidly. The reason for this success is that the model is compatible with the competitive paradigm. Or, as pointed out by Krugman (1995, 54),

> Economists understood why economic activity spreads out, not why it becomes concentrated – and thus the central model of spatial economics became one that deals only with the way competition for land drives economic activities away from a central market.

1.4.1 Space and the Competitive Paradigm

More than half a century ago, when Isard (1949) critically discussed general equilibrium analysis, he was mainly concerned with Hicks's *Value and Capital* published in 1939. Isard concluded that Hicks confined himself to "a wonderland of no dimension." He further elaborated this point on page 477 in which he recorded a conversation he had with Schumpeter, who defended the Hicksian analysis, maintaining that "transport cost is implicitly contained in production cost, and thus Hicksian analysis is sufficiently comprehensive." Isard's point was that

> production theory ... cannot justifiably treat certain production costs explicitly and other important ones implicitly in order to avoid the obstacles to analysis which the latter present. For a balanced treatment, the particular effects of transport and spatial costs in separating producers from each other must be considered.[17] They are too vital to be sidestepped through implicit treatment, as Hicks and others may be interpreted as having done.

We believe that Isard was right.

In fact, the debate about whether or not the general equilibrium model based on perfect competition is comprehensive enough to fully reflect the working of the spatial economy has a long history. On one side, general equilibrium theorists have maintained that the problem of space can be handled by defining each commodity by its physical characteristics as well as by the place (period) in which it is made available, and hence, once we have thus indexed commodities, we can essentially forget space (and time) in economic theory. This is the way Arrow and Debreu (1954) treated space (and time) in their seminal article.

On the other side, from the standpoint of the alternative view, supported by Lösch, Isard, and several others, the problem is not that simple. To capture the essential impact of space on the distribution of economic activities, new models are needed that are fundamentally different from those found in standard general equilibrium. In particular, Koopmans claimed in his *Three Essays on the State of Economic Science* that the vital effects of space become evident when our concern is the location of several economic activities and, hence, when the spatial distribution of activities itself becomes a variable. In this respect, Koopmans (1957, 154) maintained that

> without recognizing indivisibilities – in human person, in residences, plants, equipment, and in transportation – urban location problems, down to those of the smallest village, cannot be understood.

Because standard general equilibrium analysis abstains from the consideration of indivisibilities or increasing returns to scale, it will fail to capture the essential impact of transport and land when one comes to study the spatial distribution of economic activities.

In the long debate concerning the comprehensiveness of general equilibrium theory for the spatial economy, Starrett (1978) has made a fundamental contribution. The essential question is whether the competitive price mechanism is able to explain the endogenous formation of economic agglomerations. To check the ability of a spatial model to do so, the best approach is to consider the case of a homogeneous space in which economic agents are free to choose their locations. For, if any concentration of economic activities is to occur, it must be due to endogenous economic forces. Starrett has shown that if space is homogeneous and transport costly, then any competitive equilibrium is such that no transportation occurs. In other words, the economy degenerates into separated single-location groups of agents with all trades taking place within, rather than between, groups. Consequently, the perfectly competitive price mechanism alone is unable to deal simultaneously with cities and trade. This fact has a fundamental implication for the modeling of the spatial economy: If the purpose is to build a theory explaining the formation of economic agglomerations, then such a theory must depart from general competitive analysis.

Once it is recognized that the competitive equilibrium paradigm cannot be the right foundation for the space-economy, what theory is conceivable? The following is Isard's second major insight to which the alternative should be a general theory of spatial competition:

> Because of the monopoly elements which are almost invariably present in spatial relations, a broadly defined general theory of monopolistic competition can be conceived as identical with the general theory of location and space-economy (Isard 1949, 504–5).[18]

1.4.2 Spatial Competition

Ever since Sraffa (1926), it has long been recognized in the economics profession that the integration of increasing returns within the competitive model is problematic. As observed by B.C. Eaton and Lipsey (1977, 63),

> Once the firm acts as if it faces a perfectly elastic demand curve, there is nothing to restrict size from the demand side. *Size must be restricted from the cost side*. Hence, the extreme importance of eventually diminishing returns to scale in any competitive model that seeks to limit the size of plants and firms. (italics in original)

while, despite unexploited economies of scale, however, the firm size is demand constrained once consumers are dispersed across locations.

In fact, combining space and economies of scale has a profound implication for economic theory. If production involves increasing returns, a finite economy accommodates only a finite number of firms, which are *imperfect competitors*. Treading in Hotelling's footsteps, Kaldor (1935) argued that space gives this competition a particular form. Because consumers buy from the firm with the lowest price augmented by transport cost, each firm competes directly with only a few neighboring firms regardless of the total number of firms in the industry (Eaton and Lipsey 1977; Gabszewicz and Thisse 1986).

The very nature of spatial competition is, therefore, oligopolistic and should be studied within a framework of interactive decision making. This was one of the central messages conveyed by Hotelling (1929) but was ignored until economists became fully aware of the power of game theory for studying competition in modern market economies.[19] Following the application of game theory to industrial organization in the late 1970s, it became natural to study the implications of space for competition. New tools and concepts are now available to revisit and formalize the questions raised by early location theorists.

But this is not yet the end of the story. Most of the contributions to location theory by industrial organization deal with partial equilibrium models. Although a comprehensive general equilibrium model with imperfect competition has so far been out of reach and is likely to remain so for a long time (Bonanno 1990), specific models have been developed that, taken together, have significantly improved our understanding of how the spatial economy works. In particular, since the 1990s, a growing number of economists have become interested in the study of location problems, and it is fair to say that some real progress has been made. This increased interest has been partially triggered by the integration of national economies within trading blocks, such as the European Union or NAFTA, that leads to the fading of national borders. In the same vein, the study of the microeconomic underpinnings of economic development has led several economists to investigate the connection between growth and cities.[20]

1.5 PLAN OF THE BOOK

To a large extent, the organization of this book reflects what we have said in the foregoing sections. Although we have tried to make each chapter more or less self-contained, the reader may benefit from "agglomeration economies" in the course of study. Thus, the book has been organized into four parts. The first one deals with the fundamentals of geographical economics. After showing the insufficiency of the competitive paradigm for studying economic geography, we consider different issues such as the land rent formation, the structure of competition between geographically separate firms, and the provision and financing of local public goods. The second part explains the structure of metropolitan areas and the clustering of firms selling similar products. In the third part, we shift to a different geographical scale and cope with the impact of factor mobility on the location of industry. In particular, we study the role of both technological and pecuniary externalities in the interregional distribution of firms. In the last part, we offer two syntheses of various approaches taken in this book, which also suggest new lines of research. We first study how perfect competition in the land market and monopolistic competition in the product market can be combined with the aim of explaining the emergence of cities in an otherwise homogeneous setting. We then proceed by investigating the relationship between agglomeration and growth once agents have forward-looking behavior.

Needless to say, the topics covered in this book reflect our idiosyncrasies. Hence, we owe our apologies to those who have contributed to the field but who might dislike our choice of menu.

1.5.1 Part I. Fundamentals of Geographical Economics

Chapter 2 shows the insufficiency of the competitive paradigm for the formation of economic agglomerations. Specifically, we follow Starrett and show that cities, local specialization, and trade cannot arise at the competitive equilibrium of an economy with a featureless space. This criticism, because it is internal to the model, is especially powerful. After having provided an intuitive explanation for Starrett's theorem, we discuss what could be the alternative modeling strategies that will allow us to study economic agglomerations in market economies.

Chapter 3 discusses the location of divisible activities, as directly inspired by Thünen, and demonstrates how a competitive land market works regarding the allocation of land among competing activities. Once it is assumed that centers do exist through which commodities are traded, the competitive model is applicable and yields sensible results regarding the way land is organized around these centers. (The reasons for the formation of centers are postponed to Part II.) We then consider the adaptation of the Thünian model to urban economics. The main results derived in the classical context of the monocentric city are then presented.

In Chapter 4, we move to the fundamental trade-off between increasing returns and transport costs and investigate several models illustrating the importance of this trade-off for the spatial economy. Our first task is to explain why the gathering of people within a small area is able to yield scale economies in the aggregate. We consider two microeconomic foundations for such social returns. In the first, a monopolistic, competitive, intermediate sector produces nontradable goods under scale economies at the firms' level. Increasing returns are transferred in the aggregate to the final sector that would otherwise exhibit constant returns, thus showing the importance of the urban service basis for the formation and productivity of the city production system. In the second model, both firms and workers are heterogeneous, whereas wage formation is driven by a matching process. The average quality of the match rises with the population size, and this factor suggests an explanation for the tendency of wages to be higher in large metropolitan areas than in smaller cities.

We then focus on the process of competition among communities (e.g., company towns) that form to exploit scale economies. When communities are able to capitalize land rent into their payoffs, we show that increasing returns do not prevent the decentralization of the optimal allocation. Production communities, on the contrary, do not form in the absence of increasing returns. This material allows us to present the basic elements of a theory of urban systems proposed by Henderson.

Finally, we demonstrate how the process of spatial competition develops once it is recognized that geographical separation gives firms market power over consumers located in their vicinity. If firms are able to capitalize the land rent they create by their mere existence, then they find it profitable to sell at marginal costs, making money from the land rent only. An old conjecture stated by Hotelling (1938) is then proven: When there is free entry, firms' fixed costs are just covered by the aggregate land rent.

A similar line of reasoning is used in Chapter 5 but is applied to local public goods. Our first result is the Henry George theorem, which claims that public expenditure equals aggregate land rent when the population size of a city is optimal. In the same spirit as in Chapter 4, we show how competition among land developers allows for the decentralization of the efficient allocation of public goods when agents are identical in preferences and incomes. We also consider voting as an alternative decision-making mechanism to determine the location and number of facilities supplying local public goods. It is shown that voting fosters too much public infrastructure financed through too big a public budget. Once again, allowing individuals to move and to compete for land use permits us to show that the optimum is unanimously selected by consumers through voting. All these results confirm the idea that a competitive land market is a powerful device for improving the allocation of resources.

1.5.2 Part II. The Structure of Metropolitan Areas

In Chapters 6 and 7, we deal explicitly with the formation of different types of economic agglomerations within cities. Specifically, we survey and extend the literature developed in urban economics, industrial organization, and regional science to explain either the emergence of a central business district or the clustering of firms selling similar products. In Chapter 6, our frame of reference is the existence of communication externalities. We first consider partial equilibrium models, the aim of which is to determine under which conditions similar agents (households or firms) want to congregate despite their competition for land. We show that the density around the endogenous center is not high enough from the welfare point of view because each agent accounts for the benefit received from the others but not for the benefits transmitted to others.

We then move to an explicit treatment of spatial interaction between firms and households on both land and labor markets. The urban structure turns out to be the outcome of the interplay between the intensity of face-to-face communications and the level of commuting costs. Low commuting costs foster the emergence of a single central business district, thus providing a key explanation for the monocentric city. However, dispersed or polycentric structures may emerge when higher commuting costs prevail. Typically, a multiplicity of equilibria arise, and transitions from one equilibrium to another may display catastrophic changes.

Chapter 7 focuses on imperfect competition as the main explanation for the clustering of firms within cities. Without a strictly positive markup generated by product differentiation, there would be no agglomeration in the models analyzed. We deal with the case of monopolistic and oligopolistic competition and consider mobile as well as fixed consumers. As observed by Stahl (1983), product variety is a major determinant of consumers' spatial behavior. The general message of this chapter is that low transport costs together with sufficient product differentiation push economic agents toward agglomeration. The reason is that product differentiation relaxes price competition and consequently allows firms to attract more consumers when they are clustered than when any firm chooses to stand alone.

1.5.3 Part III. Factor Mobility and Industry Location

In Chapters 8 and 9, our interest shifts from cities to the spatial distribution of industries among larger spatial entities, that is, regions or nations. Hence, land consumption is no longer an issue. This is not to deny the reality of congestion effects, but we believe that they have little to do with the imbalance between big regions. At this geographical scale, the reasons for over- or underconcentration have more to do with interindustry linkages or linkages between firms and consumers, workers, or both, through the product and labor markets. Chapters 8 and 9 can be viewed as the counterpart of Chapters 6 and 7, respectively, because externalities and imperfect competition are the corresponding engines of agglomeration in each pair of chapters. Their aim is to present "clarifying examples" enhancing our understanding of how the obstacles to the spatial mobility of goods and factors affect the economic geography. In particular, these chapters provide illustrations of what is likely the main spatial feature of modern economies, namely, the emergence of a "putty-clay" economic geography. Specifically, the recent fall in trade costs seems to allow for a great deal of flexibility in where particular activities can locate, but once spatial differences have developed, they tend to become rigid. Hence, regions that were once similar may end up having very different production structures.

Chapter 8 is devoted to the impact of technological externalities, whereas Chapter 9 is concerned with pecuniary externalities expressed through monopolistic competition. In Chapter 8, we deal with the existence of urbanization economies in an otherwise standard model of regional economics. The sole presence of such externalities suffices to upset the convergence result derived in the standard neoclassical model. We then shift to localization economies to investigate the interplay between the fall in trade costs, the cost reductions associated with the implicit cooperation arising among firms located in the same area, and the intensity of competition between firms in the domestic and foreign product markets. When trade costs keep falling, an asymmetric distribution of firms emerges gradually from the interplay between these three forces.

Chapter 9 deals with what has come to be known as the "new economic geography." What drives the formation of agglomeration here is the presence of many types of pecuniary externalities such as those created by firms or workers moving from one region to the other. We will restrict ourselves to the description of the main forces driving the core–periphery structure, namely, when preference for variety and increasing returns combine to generate economic agglomerations.[21] In doing so, we compare two alternative formulations of preferences (CES versus quadratic utilities) and of transport technologies (iceberg versus proportional costs). We also study the impact of the intermediate sector (as modeled in Chapter 4) on the spatial distribution of firms and provide a welfare analysis of the core–periphery model. Finally, we complete this chapter by extending the standard framework to deal with the issue of "history versus expectation" (Krugman).

1.5.4 Part IV. Urban Systems and Regional Growth

In the final two chapters, we show how the material developed in previous chapters may be used to address two major economic issues: the formation of urban systems and the unequal growth of regions. In Chapter 10, we combine different models studied in previous chapters in order to develop a synthetic approach whose aim is to explain how and why cities emerge as a response to population growth. For that, we graft a competitive land market associated with the agricultural sector onto the canonical core–periphery model studied in Chapter 9. A monocentric configuration arises as a spatial equilibrium when the transport cost of the agricultural good is low relative to the cost of moving the industrial goods and when the total population is small. Using the intermediate input framework developed in the previous chapter leads to a wider array of results. In particular, we show that two very distinct types of monocentric patterns may emerge according to the level of intermediate inputs' transport costs. Finally, we discuss how these models can be used to explain the regular pattern of cities suggested by central place theory when population grows continuously. In addition to their theoretical interests, the results presented in this chapter shed light on the urbanization phases that took place in the United States during the second half of the nineteenth century.

As will be seen in the course of this book, geographical economics has strong connections with several branches of modern economics, including industrial organization and urban economics but also with the new theories of growth and development. In particular, economic geography and endogenous growth theory share the same framework, using monopolistic competition, increasing returns, and spillovers. This suggests the existence of a high potential for cross-fertilization. Indeed, regional growth turns out to be a new and promising topic, although it is still in its infancy.

In Chapter 11, we deal with some of the main issues addressed in the hope of convincing the reader of the relevance of further research in this domain. The main message here seems to be that, in a world of globalization, agglomeration may well be the territorial counterpart of economic growth much in the same way as growth seems to foster inequality among individuals. However, inequalities may be accompanied by a higher level of welfare even for those living on the periphery. If such preliminary results were to be confirmed, they would have farfetched implications for the modern space-economy as well as for the design of more effective economic policies.

NOTES

1. The term *agglomeration* is less ambiguous than *concentration*, which is used to describe different economic phenomena. *Agglomeration* has been introduced in location theory by Weber ([1909] 1929). Though Weber is mainly known for his work on the location of the firm (Wesolowsky 1993), his main concern was to explain the formation of industrial clusters (Isard 1956, chap. 2).
2. In this respect, R. Martin (1999, 387) is right in his criticism of economists' proclivity to use the same models "to explain the tendency for economic activity to agglomerate at various spatial scales, from the international, through the regional, to the urban and the local."
3. Before proceeding, we would like to clarify how this book relates to two recent volumes. First, the present book differs essentially from the work by Fujita, Krugman, and Venables (1999), which focuses exclusively on monopolistic competition à la Dixit–Stiglitz. In contrast, we consider a broader range of approaches and concepts, with a special focus on cities, in order to study the foundations of the spatial economy. We also cover more broadly the economics and regional science literature that have been devoted to the location of economic activities. The two books are therefore complementary, defining the frontier of geographical economics. Our book also differs from the one edited by Huriot and Thisse (2000), which deals more with various specific urban issues (e.g., the dynamics of cities when land is not malleable) or particular aspects of the process of agglomeration (e.g., the impact of globalization on the geography of financial centers) that are not covered here. Once again, there is complementarity.
4. Throughout this book, the word *city* refers to a whole urban region; we will use *city*, *metropolitan area*, and *urban area* interchangeably.
5. Throughout this book, transportation costs are broadly defined to include all impediments caused by distance such as shipping costs per se, tariff and nontariff barriers to trade, different product standards, difficulty of communication, and cultural differences.
6. It is worth noting here that Isard and Peck (1954) tried to echo Ohlin's concern about the relevance of transport costs in trade theory. Isard (1954) also strove to provide an early justification for the gravity model, which was familiar to Tinbergen as well (1962).
7. This is what Cronon (1991) calls "first nature" by contrast to "second nature," which emerges as the outcome of human beings' actions to improve upon the first one.
8. See Razin and Sadka (1997) for a synthetic presentation of migration and trade as possible substitutes.
9. It has recently been argued that capital does not necessarily flow from rich to poor regions (Lucas 1990), whereas persistent regional wage differences seem to be frequent within modern economies (Shields and Shields 1989). In addition, the empirical evidence that per capita income would converge across countries, or even

between regions of the same country, is not conclusive. Without being complete, we should like to mention Sala-i-Martin (1996), Blanchard and Katz (1992), de la Fuente and Vives (1995), de la Fuente (1997), and Quah (1996).

10. In the 1970s, another prominent trade theorist, R.G. Lipsey, vastly contributed, with B.C. Eaton, to the development of spatial economic theory (see, e.g., Eaton and Lipsey 1977, 1997).
11. An attempt to clarify the concept of Marshallian externalities is made in Section 4.2.
12. This phenomenon is similar to that encountered in network externalities. Besides the network effect, which is an agglomeration force because consumers always prefer a larger network, it is necessary to identify another effect that plays the role of a dispersion force in order to obtain different networks (see Grilo, Shy, and Thisse 2001 for a spatial model with network externalities). Note also that the issue of standardization bears some resemblance to that of agglomeration (Arthur 1994, chaps. 2 and 4).
13. In a sense, this corresponds to a revival of ideas advocated by early development theorists who used various related concepts such as the "big push" of Rosenstein-Rodan (1943), the "growth poles" of Perroux (1955), the "circular and cumulative causation" by Myrdal (1957), and the "backward and forward linkages" by Hirschman (1958). Recent additions to this cornucopia include the "dynamic economies of scale" by Kaldor (1985), the "positive feedbacks" by Arthur (1994, chap. 1) and the "complementarities" by Matsuyama (1995).
14. The idea that cities have an optimal size is old and goes back at least to Plato, for whom the ideal city has 5,040 citizens. This number does not include women, children, slaves, and foreigners, thus making the total number of residents significantly larger (we thank Yorgos Papageorgiou for having pointed out this reference to us).
15. See section 2 of part II of *The Isolated State*, which contains the extracts of posthumous papers on location theory written by Thünen between 1826 and 1842 and edited by Hermann Schumacher in 1863. The reader is referred to Fujita (2000) for more details.
16. See Ponsard (1983) for a historical survey of spatial economic theory.
17. It is not clear what Isard meant here by "the particular effects of transport and spatial costs in separating producers from each other." But, because Isard complained in the same paper about Hicks's rejection of monopolistic competition model in favor of perfect competition, we guess that "the particular effects" include the monopolistic elements that spatial costs introduce into price theory.
18. Of course, Isard does not refer here to the Dixit–Stiglitz model of monopolistic competition but more broadly to what is now called imperfect competition.
19. In this article, Hotelling's contribution to economic theory has been fundamental in many respects. For example, Mueller (1989, 180) regards Hotelling's paper as the pioneering contribution in public choice. The idea to formulate a game on price and locations according to a two-stage procedure was also extremely ingenious and original; it precedes by several decades the work of Selten on perfect equilibrium.
20. It is worth noting that preclassical economists have stressed the role of cities in the process of development and growth (see, e.g., Lepetit 1988, chap. 3, for an overview of the main contributions before Adam Smith). In particular, those economists viewed cities not only as a combination of inputs but also as a "multiplier" that leads

to increasing returns in the aggregate. In accord with modern urban economics, preclassical economists further considered cities as economic agents having the power to make decisions. Not surprisingly, their work is connected to modern theories of growth, thus suggesting that the "new" theories of agglomeration and of endogenous growth have the same historical roots. There are here several interesting questions that should be explored by historians of economic thought.

21. Using product variety as a surrogate for urban life agrees with the early work by Cantillon (1755). According to this author, the origin of cities was to be found in the concentration of land ownership, allowing landowners to live at a distance from their estates in places where they could "enjoy agreeable society," and in an agglomeration economy related to the landowners' demand, which attracted craftsmen and merchants.

PART I

FUNDAMENTALS OF GEOGRAPHICAL ECONOMICS

2

The Breakdown of the Price Mechanism in a Spatial Economy

2.1 INTRODUCTION

As a start, it is natural to ask the following question: To what extent is the competitive paradigm useful in understanding the main features of the economic landscape described in Chapter 1? The general competitive equilibrium model is indeed the benchmark used by economists when they want to study the market properties of an economic issue. Before proceeding, we should remind the reader that the essence of this model is that all trades are impersonal: When making their production or consumption decisions, economic agents need to know the price system only, which they take as given. At a competitive equilibrium, prices provide firms and consumers with all the information they need to know to maximize their profit and their utility.

The most elegant and general model of a competitive economy is undoubtedly that developed by Kenneth Arrow, Gérard Debreu, and Lionel MacKenzie. According to this model, the economy is formed by agents (firms and households) and by commodities (goods and services). A firm is characterized by a set of production plans, each production plan describing a possible input–output relation. A household is identified by a relation of preference, by a bundle of initial resources, and by shares in firms' profits. When both consumers' preferences and firms' technologies are convex, a price system (one price per commodity), a production plan for each firm, and a consumption bundle for each household exist that satisfy the following conditions at the prevailing prices:

1. Supply equals demand for each commodity;
2. Each firm maximizes its profit subject to its production set; and
3. Each household maximizes its utility under a budget constraint defined by the value of its initial endowment and shares in firms' profits.

In other words, all markets clear while each agent chooses the most preferred action at the equilibrium prices.

In this model, a commodity is defined not only by its physical characteristics but also by the place it is made available. This implies that the same good traded at different places is treated as different economic commodities.[1] Within this framework, choosing a location is part of choosing commodities. This approach integrates spatial interdependence of markets into general equilibrium in the same way as other forms of interdependence. Thus, the Arrow–Debreu model seems to obviate the need for a theory specific to the spatial context.

Unfortunately, however, matters are not that simple. As will be seen in Section 2.3, the competitive model cannot generate economic agglomerations unless strong spatial inhomogeneities are assumed. More precisely, we follow Starrett (1978) and show that introducing a homogeneous space (in a sense that will be made precise in the following paragraphs) in the Arrow–Debreu model implies that *total transport costs in the economy must be zero at any spatial competitive equilibrium*, and thus regional specialization, cities, and trade cannot be equilibrium outcomes. In other words, the competitive model per se cannot be used as the foundation for the study of a spatial economy. This is because we are interested in identifying purely economic mechanisms leading agents to agglomerate even in a featureless space. Indeed, as argued by Hoover (1948, 3),

> Even in the absence of any initial differentiation at all, i.e., if natural resources were distributed uniformly over the globe, patterns of specialization and concentration of activities would inevitably appear in response to economic, social, and political principles.

Starrett's result has far-reaching implications for our purpose. Indeed, once it is recognized that economic agents use land, they cannot all be together at the same location. As a consequence, the only equilibrium compatible with the competitive setting and a homogeneous space involves a collection of local autarkies.[2] It is thus almost impossible to think of a spatial economy in which agents are price-takers and to derive relevant and plausible results at the same time about the distribution of economic activities over a homogeneous space.

Of course, space is not homogeneous and trade may occur because the geographic distribution of resources is nonuniform, as in the neoclassical theory of international trade. Note that trade may also arise because exchange must occur at some given places, as in the Thünian model studied in Chapter 3. Although diversity in "first nature" is obviously pertinent, the unequal distribution of resources seems weak as the only explanation for agglomeration and trade.[3] Likewise, the formation of marketplaces is to be explained rather than assumed.

At this point, it is worth discussing the major assumptions made by Arrow and Debreu (1954), as well as their successors, to demonstrate the existence of prices that simultaneously equilibrate all markets. They suppose convexity of consumers' preferences and consumption sets as well as that of firms' production sets. In addition, constant unit prices imply that consumers' budget constraints are linear. These hypotheses are restrictive in themselves but in the context of a spatial economy become literally untenable. In particular, the convexity assumptions imply that consumers (producers) want to spread their consumption (production) activity over many locations as if they were ubiquitous. Because they are not, space therefore implies that some fundamental nonconvexities arises in the general equilibrium model (Scotchmer and Thisse 1999).

Convexity of preferences is not really necessary for existence of a competitive equilibrium if there are a large number (a continuum) of consumers. Nevertheless, it is worth pointing out that convexity of preferences is contradicted by the evidence regarding consumers' choice of housing. As stressed by Mirrlees (1972), with convex preferences a consumer would purchase a small quantity of a large number of goods – in particular, a small quantity of housing in many different locations.[4] This is not what consumers do. Following a now standard approach, Grimaud and Laffont (1989) relaxed the hypothesis of convex preferences by assuming a continuum of consumers and a finite set of locations and established the existence of a competitive equilibrium in an exchange economy in which consumers choose to reside in only one place.

Because each consumer resides in a small number of places (typically one), the residential choice also implies that parts of a consumer's initial endowment, especially labor force and skills, are available only at her residence. Since goods are differentiated by their location, a consumer's endowment changes with location as well and, therefore, with her consumption bundle. Furthermore, consumers commonly organize shopping itineraries to minimize the total expenditure, including transport costs. "Trip-chaining" suggests a particular structure of substitution between retail outlets that introduces nonconvexities into the budget constraint, thus affecting demand functions in complex ways.

Convex technologies are troubling in a more fundamental way.[5] The hypothesis that production sets are convex implies that production exhibits no increasing returns to scale – whatever its scale. Fragmenting a firm's operations into smaller units at different locations does not reduce the total output available from the same given inputs whereas transport costs decline. If the distribution of natural resources is uniform, the economy is such that each person produces for her own consumption, we therefore have *backyard capitalism*. Although the number of firms is given, each firm prefers a small plant at each of many locations, which again differs from what we observe in the real world. Hence, increasing returns to scale are critical in explaining the geographical distribution of productive activities. This claim has been coined the "folk theorem of spatial economics."

The introduction of increasing returns to scale into general equilibrium models has generated much interest recently. Although these attempts are interesting, they remain largely unsatisfactory mainly because they skirt the essential question posed by Sraffa (1926, 545): To what extent is price-taking behavior compatible with increasing returns to scale? Suppose the firm size that minimizes average production cost is "large" relative to the size of the market. A price-taking equilibrium could not have "many" firms, each operating at inefficiently small scale, because each such firm would have a profit incentive to increase its output. Hence, the market can only accommodate a "few" firms of efficient size. But with only a few firms, how does one justify the hypothesis that firms treat prices as given, for firms must realize that their size permits them to influence prices to their own advantage? This takes us to the problem of market size. Although a price-taking equilibrium does not in itself require many agents, they seem necessary to justify the behavioral hypothesis that agents are price-takers (Novshek and Sonnenschein 1987).

Even if the economy is large, so that the total number of firms can be large, the geographic dispersion of consumption causes production to be dispersed and local markets to be "small." Thus, the combination of increasing returns and geographically dispersed consumption renders untenable the hypothesis that many firms compete in each market. If one returns to the suggestion of Arrow and Debreu to distinguish goods by their location, most markets are probably characterized by a small number of firms (if any) that, as a consequence, do not behave competitively.

The remainder of the chapter is organized as follows. The inadequacy of the competitive assumption for spatial economics is demonstrated by means of a simple example in Section 2.2. The robustness of the conclusions drawn from this example is then examined in Section 2.3 in which what we call the spatial impossibility theorem is proven: No competitive equilibrium involving trade across locations exists in a homogeneous space. Unlike the criticisms discussed above, the point made here is internal to the theory, which makes it stronger. A competitive equilibrium may exist, however, when space is heterogeneous. This leads us quite naturally to investigate in Section 2.4 the validity of the first welfare theorem in a spatial economy with a finite number of locations and a heterogeneous space. In Section 2.5, we explore the possibility of decentralizing the optimal configuration of firms through a nonlinear price system when there are indivisibilities at the plant level in a spatial economy with a finite number of locations and a heterogeneous space. We do this in the context of a classical operations research model known as the simple plant location problem. We again find a negative answer even though we allow for two-part tariffs. Section 2.6 presents our conclusions.

2.2 THE QUADRATIC ASSIGNMENT PROBLEM

We begin our discussion by considering the quadratic assignment problem introduced by Koopmans and Beckmann (1957). Assume that M firms are to be assigned to M locations. The *quadratic assignment problem* is defined by the following set of assumptions[6]: Each firm is indivisible, and the amount of land available at each location is such that a single firm can be set up there.

Hence, every firm must be assigned to a single location, and every location can accommodate only one firm. Each firm produces a fixed amount of goods and uses one unit of land as well as fixed amounts of the goods produced by the others. Suppose further that the technology used by each firm is not affected by the chosen location. Finally, shipping a good from a location to another location involves a positive cost.

Although the issue addressed by Koopmans and Beckmann was the possibility of sustaining the optimal assignment through a competitive price system, we can see that any feasible assignment cannot be decentralized through a competitive price system. To illustrate the nature of the difficulties encountered, we restrict ourselves to the case of two firms, denoted $i = 1, 2$, and two locations, denoted $r = A, B$.[7] Without loss of generality, we assume that firm 1 is assigned to A and firm 2 to B. Firm i produces q_i units of good i and purchases q_j units of good j from the other firm j regardless of its own location. Firm i also receives a revenue $a_i > 0$ from other activities with the rest of the world, which does not depend on its location. Finally, each good i can be shipped from its place of production to the other location by a carrier at a given cost $t_i > 0$.

To study the sustainability of this assignment, we follow the suggestion of Arrow and Debreu by considering the same good at locations 1 and 2 as two different commodities, each with its own price. Let p_{ir} be the price of good i at location r and R_r be the rent to be paid by a firm for using one unit of land at location r. Firm 1's profit in location A is defined as follows:

$$\pi_{1A} = a_1 + p_{1A}q_1 - p_{2A}q_2 - R_A.$$

A similar expression holds for firm 2 at location B. If this price system sustains the foregoing configuration, then, as shown by Samuelson (1952), the equilibrium prices p_{ir} must satisfy the following conditions:

$$p_{1B} = p_{1A} + t_1 > p_{1A} \tag{2.1}$$

$$p_{2A} = p_{2B} + t_2 > p_{2B}. \tag{2.2}$$

In other words, the price of good 1 (2) in location B (A) is equal to its price in location A (B) plus the corresponding transport cost $t_1(t_2)$. Hence, total profits are equal to

$$\pi_{1A} + \pi_{2B} = a_1 + a_2 - (t_1q_1 + t_2q_2 + R_A + R_B),$$

which are positive if a_1 and a_2 are sufficiently large.

We now show that it is impossible to find values for the rents R_A and R_B such that both firms 1 and 2 maximize their own profit at locations A and B, respectively. Without loss of generality, assume that $R_A \geq R_B$. Then, if firms behave competitively, one can readily verify that firm 1 would earn a strictly higher profit by setting up at location B. Indeed, if firm 1 sets up at location B, its profit is

$$\pi_{1B} = a_1 + p_{1B}q_1 - p_{2B}q_2 - R_B.$$

Using (2.1) and (2.2), we can then readily verify that

$$\pi_{1B} - \pi_{1A} = t_1q_1 + t_2q_2 + R_A - R_B > 0. \tag{2.3}$$

Hence, firm 1 always has an incentive to move. This reasoning can easily be extended to the case of M firms and locations. See chapter appendix for a proof. In other words, when locations have identical exogenous attributes, no feasible location pattern of firms can be sustained as a competitive equilibrium in the quadratic assignment problem. In deriving this surprising conclusion, it has been assumed that firms believe that "changing place" does not affect the prevailing prices of goods and land rents. This assumption, however, is the essence of a price-taking equilibrium.

The reader might believe that this negative result is an artifact of the two-location, two-firm setting. For example, consider three firms 1, 2, and 3 located respectively at A, B, and C, which are colinear. Firms 2 and 3 use good 1 produced by firm 1, and firm 3 also uses good 2 produced by firm 2. Let $t_i\ (r, s)$ be the transport cost per unit of good i between r and s. Then, provided that $t_i(A, C) > t_i(A, B)$, the price system supporting such a configuration must satisfy the following relations:

$$p_{1C} = p_{1A} + t_1(A, C) > p_{1A} + t_1(A, B) = p_{1B}$$
$$p_{2C} = p_{2B} + t_2(B, C) > p_{2B}.$$

In this case, if firm 2 moves to location C, it can sell its output there at a higher price but must also pay a higher price for input 1. Hence, if p_{1C} is sufficiently larger than p_{1B}, firm 2 may find it unprofitable to relocate to C, thus suggesting that a competitive equilibrium exists. Yet, the nonexistence of an equilibrium does carry over to the quadratic assignment problem with an arbitrary number of firms and locations when locations are a priori equally attractive from the viewpoint of firms. The proof is presented in the appendix to this chapter.

One might also think that the result is caused by the specifics of the quadratic assignment problem. In the next section, we show that such a breakdown of the competitive price mechanism holds for a general spatial economy.

2.3 THE SPATIAL IMPOSSIBILITY THEOREM

To gain more insights about the proof of the main result of this section, it is desirable to go one step further in the preceding example by computing firm 2's incentive to move as in (2.3), that is,

$$\pi_{2A} - \pi_{2B} = t_2q_2 + t_1q_1 + R_B - R_A. \tag{2.4}$$

Summing expressions (2.3) and (2.4) yields $2(t_1q_1 + t_2q_2)$. This means that the sum of firms' incentives to move is equal to twice the total transport costs. Because this sum must be nonpositive for a competitive equilibrium to exist, transportation cannot occur in such an equilibrium. This suggests that competitive pricing and positive transport costs are incompatible in a homogeneous spatial economy, for the incentive to change location is of the same order of magnitude as transport costs.

Our objective in this section is to show that this property, which we call the spatial impossibility theorem, holds in a general setting. To facilitate comparison, we use standard notation from general equilibrium theory. However, to make our point more transparent, we explicitly distinguish prices and goods by their location and separate land as well as transport services from the remaining goods.

2.3.1 Competitive Equilibrium in a Homogeneous Spatial Economy

Consider a spatial economy formed by two regions A and B that can both accommodate a large number of firms and households. Each region $r = A, B$ is endowed with the same positive amount of land S. There are n goods (excluding land and transport services), and each of them can be moved from one region to the other by using transport services. There are M firms and N households; to ease the burden of notation, M and N also denote the sets of firms and households. By definition of a static model, firms and households are located "nowhere" before choosing their equilibrium place so that they can choose at zero cost the region in which they want to conduct their activities. When firm $f \in M$ sets up in region $r = A, B$, a production plan for this firm is given by a vector $\boldsymbol{y}_{fr}$ of n goods (outputs are positive and inputs are negative) and by a positive amount of land s_{fr} in region r. The firm's production set is denoted by $Y_{fr} \subset R^{n+1}$; this set may vary with the region in which the firm is established. Household $h \in N$ resides and works in the same region $r = A, B$, and its consumption plan is given by a vector $\boldsymbol{x}_{hr}$ of n goods (a positive component means that the household has a positive demand for the good, whereas a negative component means that the household is a supplier of the good – such as labor) and by a positive amount of land s_{hr} in region r. The household's consumption set is given by $X_{hr} \subset R^{n+1}$. Household h has a utility function U_{hr} defined over X_{hr}, which may both change with the region in which the household is located, together with an initial endowment of goods ω_h and a land endowment $\tilde{\boldsymbol{s}}_h = (\tilde{s}_{hA}, \tilde{s}_{hB})$. Because we consider location a separate attribute, we may assume that the same endowment in goods (e.g., labor) is available in any region in which the consumer resides and works; by contrast, the land endowment is immobile.[8]

Transportation within each region is costless, but shipping goods from one region to the other requires resources. Without loss of generality, transportation between the two regions is accomplished by a profit-maximizing carrier (or broker) who purchases goods in a region at the market prices prevailing in this region and sells them in the other region at the corresponding market prices while using goods and land in each region as inputs. The carrier ships an (nonnegative) export plan $\boldsymbol{E}_{AB} \in R^n$ of goods from A to B and an (nonnegative) export plan $\boldsymbol{E}_{BA} \in R^n$ from B to A using (nonpositive) vectors $\boldsymbol{y}_{tr} \in R^n$ of inputs and nonnegative amounts of land s_{tr} bought in both regions A and B. The set of feasible transportation plans for the carrier is denoted by $Z_t \subset R^{4n+2}$.

Let M_r be the set of firms and N_r the set of households located in region $r = A, B$ so that $M = M_A \cup M_B$ and $N = N_A \cup N_B$. An *allocation* is defined by the set N_r of households residing in region $r = A, B$, by the set M_r of firms located in region $r = A, B$, by N consumption plans $(\boldsymbol{x}_{hr}, s_{hr})$, by M production plans $(\boldsymbol{y}_{fr}, s_{fr})$, and by two export plans $\boldsymbol{E}_{AB}, \boldsymbol{E}_{BA}$ together with the associated input vectors $\boldsymbol{y}_{tA}, \boldsymbol{y}_{tB}$ and land requirements s_{tA}, s_{tB}. Therefore, an allocation describes both the location and the consumption or production activities of each household or firm as well as the transportation activity carried on by the carrier.

For an allocation to be feasible, the following material balance conditions must be met:

1. for goods in region A

$$\sum_{h \in N_A} \boldsymbol{x}_{hA} + \boldsymbol{E}_{AB} - \boldsymbol{y}_{tA} = \sum_{h \in N_A} \boldsymbol{\omega}_h + \sum_{f \in M_A} \boldsymbol{y}_{fA} + \boldsymbol{E}_{BA} \tag{2.5}$$

2. for goods in region B

$$\sum_{h \in N_B} \boldsymbol{x}_{hB} + \boldsymbol{E}_{BA} - \boldsymbol{y}_{tB} = \sum_{h \in N_B} \boldsymbol{\omega}_h + \sum_{f \in M_B} \boldsymbol{y}_{fB} + \boldsymbol{E}_{AB} \tag{2.6}$$

3. for land in region $r = A, B$

$$\sum_{h \in N_r} s_{hr} + \sum_{f \in M_r} s_{fr} + s_{tr} \leq \sum_{h \in N_r} \tilde{s}_{hr} \equiv S, \tag{2.7}$$

where $(\boldsymbol{x}_{hr}, s_{hr}) \in X_{hr}$, $(\boldsymbol{y}_{fr}, s_{fr}) \in Y_{fr}$ and $(\boldsymbol{E}_{AB}, \boldsymbol{E}_{BA}, \boldsymbol{y}_{tA}, \boldsymbol{y}_{tB}, s_{tA}, s_{tB}) \in Z_t$.[9]

Finally, a *competitive equilibrium* for the economy described earlier in this section is given by a price system – that is, two vectors $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ for the goods and a land rent pattern (R_A, R_B) – and a feasible allocation as above, such that:

1. all markets clear in each region r; that is, (2.5)–(2.7) hold;
2. each firm $f \in M_r$ maximizes its profit at the chosen location and feasible production plan

$$\pi_{fr} \equiv \boldsymbol{p}_r \cdot \boldsymbol{y}_{fr} - R_r s_{fr} \geq \boldsymbol{p}_s \cdot \hat{\boldsymbol{y}}_{fs} - R_s \hat{s}_{fs}$$

for all $(\hat{\boldsymbol{y}}_{fs}, \hat{s}_{fs}) \in Y_{fs}$ and $s = A, B$;
3. each household $h \in N_r$ maximizes its utility at the chosen location and

consumption plan subject to the household's budget constraint:

$$U_{hr}(\boldsymbol{x}_{hr}, s_{hr}) \geq U_{hs}(\hat{\boldsymbol{x}}_{hs}, \hat{s}_{hs})$$

for all $(\hat{\boldsymbol{x}}_{hs}, \hat{s}_{hs}) \in X_{hs}$ and $s = A, B$ such that

$$\begin{aligned}\boldsymbol{p}_s \cdot \hat{\boldsymbol{x}}_{hs} + R_s \hat{s}_{hs} \leq \boldsymbol{p}_s \cdot \boldsymbol{\omega}_h + \sum_{r\in\{A,B\}} R_r \tilde{s}_{hr} \\ + \sum_{r\in\{A,B\}} \sum_{f\in M_r} \theta_{hf}\pi_{fr} + \theta_{ht}\pi_t,\end{aligned}$$

where θ_{hf} is the share of consumer h in firm f's profits and θ_{ht} the consumer's share in the carrier's profit π_t; and

4. the carrier maximizes its profit defined by

$$\begin{aligned}\pi_t = (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot \boldsymbol{E}_{AB} + (\boldsymbol{p}_A - \boldsymbol{p}_B) \cdot \boldsymbol{E}_{BA} + \boldsymbol{p}_A \cdot \boldsymbol{y}_{tA} \\ + \boldsymbol{p}_B \cdot \boldsymbol{y}_{tB} - R_A s_{tA} - R_B s_{tB}\end{aligned} \quad (2.8)$$

subject to its transportation plan being in $Z_t \subset R^{4n+2}$.

Space is said to be *homogeneous* when (1) the utility function U_h and the consumption set X_h are the same regardless of the region where household h resides, and (2) the production set Y_f is independent of the region elected by firm f. In other words, consumers and producers have no intrinsic preferences for one region over the other.

Suppose that space is homogeneous. The profit of firm f located in region A is given by the following expression:

$$\pi_{fA} = \boldsymbol{p}_A \cdot \boldsymbol{y}_{fA} - R_A s_{fA}.$$

Space being homogeneous, the production plan $(\boldsymbol{y}_{fA}, s_{fA})$ is also possible in region B. If firm f were to locate in region B while keeping the same production plan, its profit would become:

$$\pi_{fB} = \boldsymbol{p}_B \cdot \boldsymbol{y}_{fA} - R_B s_{fA}.$$

Hence, for firm f the incentive to move from A to B is defined by the difference in profit earned in each of the two regions:[10]

$$I_f(A, B) = \pi_{fB} - \pi_{fA} = (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot \boldsymbol{y}_{fA} - (R_B - R_A) s_{fA}. \quad (2.9)$$

Clearly, an expression similar to (2.9) holds for every firm set up in region B.

Consider now a household h residing in region A. The household's residual income (neglecting its share in firms' profit and income from land, which are independent of the place of residence) is defined by the expression

$$B_{hA} = \boldsymbol{p}_A \cdot (\boldsymbol{\omega}_h - \boldsymbol{x}_{hA}) - R_A s_{hA}.$$

If this consumer were to locate in region B while keeping the same consumption plan, she would derive the same utility from $(\boldsymbol{x}_{hA}, s_{hA})$, and thus only the consumer's residual income in region B would matter:

$$B_{hB} = \boldsymbol{p}_B \cdot (\boldsymbol{\omega}_h - \boldsymbol{x}_{hA}) - R_B s_{hA}.$$

Hence, if there is no satiation, the consumer's incentive to move from A to B is given by the difference in her residual income in each of the two regions:

$$I_h(A, B) = B_{hB} - B_{hA} = (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot (\boldsymbol{\omega}_h - \boldsymbol{x}_{hA}) - (R_B - R_A)s_{hA}. \tag{2.10}$$

Again, an expression similar to (2.10) holds for a household residing in B.

Summing (2.9) and (2.10) for all firms and households across the two regions, we obtain:

$$\begin{aligned} I = {} & (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot \left(\sum_{f \in M_A} \boldsymbol{y}_{fA} + \sum_{h \in N_A} (\boldsymbol{\omega}_h - \boldsymbol{x}_{hA}) \right) \\ & + (\boldsymbol{p}_A - \boldsymbol{p}_B) \cdot \left(\sum_{f \in M_B} \boldsymbol{y}_{fB} + \sum_{h \in N_B} (\boldsymbol{\omega}_h - \boldsymbol{x}_{hB}) \right) \\ & - (R_B - R_A) \left(\sum_{f \in M_A} s_{fA} + \sum_{h \in N_A} s_{hA} \right) \\ & - (R_A - R_B) \left(\sum_{f \in M_B} s_{fB} + \sum_{h \in N_B} s_{hB} \right). \end{aligned}$$

Using the material balance conditions (2.5)–(2.7), we can rewrite this expression as follows:

$$\begin{aligned} I = {} & (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot (\boldsymbol{E}_{AB} - \boldsymbol{E}_{BA} - \boldsymbol{y}_{tA}) + (\boldsymbol{p}_A - \boldsymbol{p}_B) \cdot (\boldsymbol{E}_{BA} - \boldsymbol{E}_{AB} - \boldsymbol{y}_{tB}) \\ & + (R_B - R_A)(s_{tA} + \phi_A - S) + (R_A - R_B)(s_{tB} + \phi_B - S), \end{aligned}$$

where ϕ_r is the amount of land unused in region $r = A, B$. Adding $(\boldsymbol{p}_A + \boldsymbol{p}_B) \cdot (\boldsymbol{y}_{tA} + \boldsymbol{y}_{tB}) - 2(R_A s_{tA} + R_B s_{tB})$ to the first two terms, subtracting the same expression from the last two terms, and regrouping them in the resulting expression yields

$$\begin{aligned} I = 2\bigg[& \pi_t + \frac{\boldsymbol{p}_A + \boldsymbol{p}_B}{2} \cdot (-\boldsymbol{y}_{tA} - \boldsymbol{y}_{tB}) + \frac{R_A + R_B}{2}(s_{tA} + s_{tB}) \\ & + \frac{R_A + R_B}{2}(\phi_A + \phi_B) \bigg] \end{aligned} \tag{2.11}$$

in which we have used (2.8) as well as $R_A \phi_A = R_B \phi_B = 0$ because the land rent is zero when all the land available in a region is not used.

In words, (2.11) means that the aggregate incentives for firms and households to move from one region to the other is equal to twice the carrier's profit

$(2\pi_t)$ plus two terms that can be interpreted as twice the carrier's costs evaluated at the average prices $(\boldsymbol{p}_A + \boldsymbol{p}_B)/2$ and rent $(R_A + R_B)/2$ (they are called pseudo-transport costs), plus twice the value of vacant land evaluated at the average rent $(R_A + R_B)/2$. Because the carrier maximizes its profit, π_t cannot be negative. In addition, unless the equilibrium involves no transportation, the pseudo-transport costs are strictly positive because shipping goods requires scarce resources: some components of the vectors $\boldsymbol{y}_{tr}$ must be negative, whereas no component is positive by assumption; similarly the quantities of land used must be nonnegative, whereas the land rents are nonnegative. Clearly, the last term in (2.11), $(R_A + R_B)(\phi_A + \phi_B)$, cannot be negative, for land rents are nonnegative. Consequently, the global incentives to move are always strictly positive for any allocation involving costly trade between the two regions.

In a competive equilibrium, no agent has a positive incentive to move. Therefore, we may conclude with the following theorem:

The Spatial Impossibility Theorem. *Assume a two-region economy with a finite number of consumers and firms. If space is homogeneous, transport is costly, and preferences are locally nonsatiated, there is no competitive equilibrium involving transportation.*

What does this theorem mean? If economic activities are perfectly divisible, a competitive equilibrium exists and is such that each location operates as an autarky. For example, when households are identical, regions have the same relative prices and the same production structure (backyard capitalism). This is hardly a surprising outcome because, by assumption, there is no reason for economic agents to distinguish among locations and each activity can operate at an arbitrarily small level. Firms and households thus succeed in reducing transport costs at their absolute minimum, namely zero.

However, as observed by Starrett (1978, 27), when economic activities are *not* perfectly divisible, the transport of goods or people between some places becomes unavoidable:

> as long as there are some indivisibilities in the system (so that individual operations must take up space) then a sufficiently complicated set of interrelated activities will generate transport costs.

In this case, the spatial impossibility theorem tells us that no competitive equilibrium exists.

This is clearly a surprising result that requires more explanation. When both regions are not in autarky, one should keep in mind that the price system must perform two different jobs simultaneously: (1) to support trade between regions (while clearing the markets in each region), and (2) to prevent firms and households from relocating. The spatial impossibility theorem says that, in the case of a homogeneous space, it is impossible to hit two birds with one stone: the price gradients supporting trade bear wrong signals from the viewpoint of locational stability. Indeed, if a set of goods is exported from A to B, then the associated positive price gradients induce producers located in region A (who seek a higher revenue) to relocate in region B, whereas region B's buyers (who seek lower prices) want to relocate in A. Likewise, the export of another set of goods from B to A encourages such "cross-relocation." The land rent differential between the two regions can discourage the relocation in one direction only. Hence, as long as trade occurs at positive costs, some agents always want to relocate.[11]

To ascertain the fundamental cause for this nonexistence, it is helpful to illustrate the difficulty encountered by using a standard diagram approach. Depicting the whole trade pattern between two regions would require a diagram with six dimensions (two tradable goods and land at each location), which is a task beyond our capability. We thus focus on a suballocation formed by the feasible trade patterns of good i between A and B only and keep the other elements fixed. Because the availability of the same physical good at two distinct locations gives rise to two different commodities, this is equivalent to studying a standard transformation between two different economic goods.

Suppose that one unit of good i is produced by one firm at either location by using a fixed bundle of inputs. For simplicity, the cost of these inputs is assumed to be the same in both locations. The good is shipped according to an iceberg technology (Samuelson 1954a): when x_i units of the good are moved between A and B, only a fraction x_i/Υ arrives at destination, where $\Upsilon > 1$, whereas the rest melts away en route. In this context, if the firm is located in A, then the output is represented by point E on the vertical axis in Figure 2.1; if

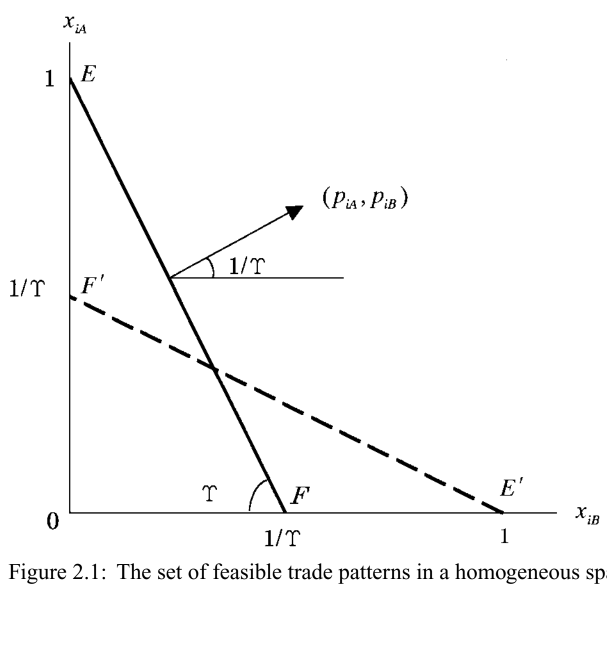


Figure 2.1: The set of feasible trade patterns in a homogeneous space.

the entire output is shipped to B, then the fraction $1/\Upsilon$ arrives at B, which is denoted by point F on the horizontal axis. Hence, when the firm is at A, the set of feasible allocations of the output between the two locations is given by the triangle OEF. Likewise, if the firm locates at B, the set of feasible allocations between the two places is now given by the triangle $OE'F'$. Hence, when the firm is not located, the set of feasible allocations is given by the union of the two triangles.

Let the firm be set up at A and assume that the demand conditions are such that good i is consumed in both locations so that trade occurs. To support any point belonging to the frontier EF, the price vector (p_{iA}, p_{iB}) must be such that $p_{iA}/p_{iB} = 1/\Upsilon$, as shown in Figure 2.1. However, under these prices, it is clear that the firm can obtain a strictly higher profit by locating in B and choosing the production plan E' in Figure 2.1. This implies that there is no competitive price system that can support both trade and a profit-maximizing location for the firm.

This difficulty arises from the nonconvexity of the set of feasible allocations. If transportation were costless, this set would be given by the triangle OEE' in Figure 2.1, which is convex. In this case, the firm would face no incentive to relocate. Similarly, if the firm's production activity were perfectly divisible, the set of feasible allocations would again be equal to the triangle OEE', and no difficulty would arise.

Therefore, we may conclude that *the fundamental reason for the spatial impossibility theorem is the nonconvexity of the set of feasible allocations caused by positive trade costs and the fact that agents have an address in space,* even though the individual land consumption is endogenous.

2.3.2 Examples

To investigate the implications and meaning of the spatial impossibility theorem further, we find it illustrative to study some examples involving heterogeneous groups of agents with strong intragroup and weak intergroup linkages.

Example 2.1 Consider an economy formed by two groups of agents such that the members of each group have strong internal linkages, whereas each group produces a distinct set of final goods indispensable to both groups. When the transport costs of the final goods are not too high, it is then natural to expect each group of agents to agglomerate into a separate region and to trade their net outputs across regions. We show in this example that if space is homogeneous (in the sense defined in page 33), then there is no competitive price system that can support such a "natural" spatial configuration. We also show that locating both groups within one region (on the assumption that this is physically possible) is not either equilibrium when the marginal utility of land is positive for at least one agent.

Consider an economy with two locations A and B, two firms 1 and 2, and two workers a and b. The two locations have the same amount of land S. Worker

$h = a$ (b) is endowed with one unit of labor of type $i = 1$ (2) together with one-half of the land available at each location and half of the profits earned by the two firms. Both workers have the same utility function,

$$U = x_1^{\beta/2} x_2^{\beta/2} s^{1-\beta} \quad 0 < \beta < 1, \tag{2.12}$$

where x_i is the consumption of good $i = 1, 2$ and s is the land consumption. Firm f (= 1, 2) produces good $i = f$ at a single location using two inputs: a fixed amount $\bar{s}$ of land and a variable amount l_i of type of labor i acquired at the same location. Its production function is given by

$$Q_i = l_i^{\alpha} \quad 0 < \alpha < 1. \tag{2.13}$$

Each good is transported according to the same iceberg technology.

Let p_{ir}, w_{ir}, and R_r be respectively the price of good i, the wage rate of the type of labor i, and the land rent at location r. If firm i locates at r, its profit is as follows:

$$\pi_{ir} = p_{ir} Q_i - w_{ir} l_i - R_r \bar{s}.$$

If this firm behaves competitively, maximizing π_{ir} yields

$$\alpha p_{ir} l_i^{\alpha - 1} = w_{ir}. \tag{2.14}$$

Let Y_r be the total income of a worker residing at r. Then, using (2.12), we obtain the demands of such a worker as follows:

$$x_{ir} = \frac{\beta}{2} \frac{Y_r}{p_{ir}} \tag{2.15}$$

$$s_r = (1 - \beta) \frac{Y_r}{R_r}. \tag{2.16}$$

The land constraint $S = s_r + \bar{s}$ at r together with (2.16) yields

$$R_r = \frac{(1 - \beta) Y_r}{S - \bar{s}}. \tag{2.17}$$

Because firm f (= 1, 2) uses only the type of labor $i = f$, it is reasonable to expect firm f to locate with the worker of type i in equilibrium. Hence, since the two locations are identical, there are only two possible candidates as equilibrium configurations:

1. Dispersion: firm 1 and worker a locate together in A, whereas firm 2 and worker b are in B;
2. Agglomeration: both firms and both workers locate together in, say, A.

We now consider each configuration in turn and examine whether it can be supported by a competitive price system. First, let us address the dispersed configuration. To begin, we fix locations of all agents as they are in this configuration and show the existence and uniqueness of a competitive equilibrium.

Because the whole setting is symmetric, in equilibrium it follows that

$$w_{1A} = w_{2B} = 1, \tag{2.18}$$

where labor is chosen as the numéraire. Because the whole amount of labor is used in equilibrium, we may set $l_i = 1$ in (2.14) so that, using (2.18), we have

$$p_{1A} = p_{2B} = 1/\alpha \tag{2.19}$$

and, hence, from the iceberg transport technology with parameter $\Upsilon > 1$

$$p_{1B} = p_{2A} = \Upsilon/\alpha. \tag{2.20}$$

Setting $l_i = 1$ in (2.13), we also obtain

$$Q_1 = Q_2 = 1. \tag{2.21}$$

The total demand for good 1, say, is

$$D_1 = x_{1A} + \Upsilon x_{1B},$$

where the last term represents the quantity of good 1 to be shipped from A for the quantity x_{1B} to be available for consumption in B. Substituting (2.15) into this equation and using (2.19) and (2.21), we get

$$D_1 = D_2 = \alpha\beta Y,$$

where Y is the income common to each location. Then, by equating supply and demand ($D_i = Q_i$) and using (2.14)–(2.21), we obtain

$$\begin{aligned} R_A = R_B &= \frac{1}{\alpha\beta}\frac{1-\beta}{S-\bar{s}} \\ \pi_{1A} = \pi_{2B} &= \frac{1}{\alpha\beta}\frac{\beta S-\bar{s}}{S-\bar{s}} - 1 \\ x_{1A} = x_{2B} &= \frac{1}{2} \\ x_{1B} = x_{2A} &= \frac{1}{2\Upsilon}. \end{aligned} \tag{2.22}$$

Each firm's equilibrium profits are nonnegative if and only if

$$\alpha \leq \frac{S-\bar{s}/\beta}{S-\bar{s}}. \tag{2.23}$$

Therefore, if this condition holds, a unique market equilibrium exists under the locations corresponding to dispersion. Condition (2.23) is satisfied when each firm's land requirement is sufficiently small compared with S.

Let us evaluate the incentive to move of each agent in A. In doing so, we first note that w_{1B} (wage rate for type 1 at B) is not defined. In fact, we show next that, regardless of its values, at least one agent in region A has an incentive to move. To show this, let us set the value of w_{1B} at an arbitrary level. Then, the incentive for firm 1 to move from A to B (defined under the same production plan) is such that

$$\begin{aligned} I_1(A, B) &= \pi_{1B} - \pi_{1A} \\ &= (p_{1B} - p_{1A})Q_1 - (w_{1B} - w_{1A}) \end{aligned}$$

because $R_A = R_B$. Similarly, the incentive for worker a to move from A to B (while keeping the same consumption plan) is defined as

$$\begin{aligned} I_a(A, B) &= B_{1B} - B_{1A} \\ &= (w_{1B} - p_{1B}x_{1A} - p_{2B}x_{2A} - R_B s_A) \\ &\quad - (w_{1A} - p_{2A}x_{1A} - p_{2A}x_{2A} - R_A s_A) \\ &= (w_{1B} - w_{1A}) - (p_{1B} - p_{1A})x_{1A} - (p_{2B} - p_{2A})x_{2A}. \end{aligned}$$

Summing these two expressions, we obtain

$$\begin{aligned} I(A, B) &= (p_{1B} - p_{1A})(Q_1 - x_{1A}) - (p_{2B} - p_{2A})x_{2A} \\ &= (p_{1B} - p_{1A})\Upsilon x_{1B} - (p_{2B} - p_{2A})x_{2A} \\ &= p_{1A}(\Upsilon - 1)\Upsilon x_{1B} + p_{2B}(\Upsilon - 1)x_{2A} > 0 \end{aligned}$$

in which we have used (2.20). Therefore, at least firm 1 or worker a has an incentive to move from A to B because $\Upsilon > 1$, thus implying that dispersion is not a competitive equilibrium once agents can freely choose their location.

Consider now the agglomeration in which both firms and workers locate together in A. Using the equilibrium conditions (2.14) to (2.17), one can readily verify that the corresponding equilibrium prices are as follows:

$$\begin{aligned} & w_{1A} = w_{2A} = 1 \\ & p_{1A} = p_{2A} = 1/\alpha \\ & Q_1 = x_{1A} = Q_2 = x_{2A} = 1 \\ & R_A = \frac{2\alpha\beta(1-\beta)}{S - \bar{s}} \quad \text{and} \quad R_B = 0 \\ & \pi_{1A} = \pi_{2A} = \frac{1}{\alpha\beta}\frac{\beta S - \bar{s}}{S - \bar{s}} - 1. \end{aligned} \tag{2.24}$$

Because (2.22) and (2.24) imply that the profit of each firm is the same in both configurations, a competitive equilibrium exists when (2.23) holds. Because all agents are located in A, they do not trade with B. Hence, given any equilibrium wage rates and prices at B, when all agents are in B all terms related to transactions between them cancel out, and thus the aggregate incentive for all agents to move from A to B is equal to the land cost saving:

$$\begin{aligned} I(A, B) &= (R_A - R_B)S \\ &= R_A S > 0. \end{aligned}$$

Accordingly, agglomeration is not a spatial equilibrium, either.

Example 2.2 To illustrate the role of externalities, we now consider a setting similar to the one explored in Example 2.1 in which we allow for M replicas of each of four agents, where M is assumed to be an integer number. The $2M$ workers equally share the profit of each firm as well as the land rent at each location. When M is an even number, a competitive equilibrium always exists in which $M/2$ firms and workers of each type are located in A and B, in which case there is autarky. This is not a very exciting outcome.

To obtain a competitive equilibrium with trade, we change the assumptions of the replicated economy as follows. First, for simplicity, we assume that the $2M$ workers are equally productive in either type of firms. Second, and this is more fundamental, the production function (2.13) of each firm of type i is now as follows:

$$Q_{ir} = a(M_{ir})l^{\alpha} \quad i = 1, 2 \text{ and } r = A, B, \tag{2.25}$$

where l is the amount of labor used by a firm and $a(M_{ir})$ is an increasing function of the number (M_{ir}) of type i firms at r. The function $a(M_{ir})$ represents the so-called Marshallian externality associated with the local agglomeration of firms belonging to the same industry.

Consider the dispersed configuration in which M firms of type 1 and M workers locate together in A, whereas M firms of type 2 and M workers are at B, and find under which conditions this configuration is an equilibrium. Because each firm uses one unit of labor, the two locations are formally symmetric. Hence, as in the foregoing, we can readily obtain the following equilibrium values:

$$w_A = w_B \equiv 1 \tag{2.26}$$

$$\begin{aligned} p_{1A} &= p_{2B} = 1/[\alpha a(M)] \\ p_{1B} &= p_{2A} = \Upsilon/[\alpha a(M)] \qquad (2.27) \\ Q_{1A} &= Q_{2B} = a(M) \\ R_A &= R_B = \frac{M}{\alpha\beta}\frac{1-\beta}{S - M\bar{s}} \\ \pi_{1A} &= \pi_{2B} = \frac{1}{\alpha\beta}\frac{\beta S - M\bar{s}}{S - M\bar{s}} - 1. \end{aligned}$$

When the lot size $\bar{s}$ of each firm is sufficiently small, the profit of each firm is nonnegative, and thus a unique market equilibrium exists under the dispersed configuration.

Let us examine whether an agent has an incentive to relocate. All workers attain the same utility level and, therefore, have no reason to move. In considering the possible relocation of a firm, we assume that the number M is so large that each firm takes the externality levels, $a(M)$ and $a(0)$, as given when making its locational decision. In such a context, if a type 1 firm, say, moves from A to B, then its profit is

$$\pi_{1B} = p_{1B} a(0) l^{\alpha} - w_B l - R_B \bar{s}.$$

Using (2.26) and (2.27), the profit-maximizing labor input l_{1B} at B is equal to

$$(l_{1B})^{\alpha-1} = \frac{1}{\Upsilon} \frac{a(M)}{a(0)}, \tag{2.28}$$

which yields

$$\pi_{1B} = \frac{1-\alpha}{\alpha} l_{1B} - R_B \bar{s}.$$

Because

$$\pi_{1A} = p_{1A} a(M) - w_A - R_A \bar{s} = (1/\alpha) - 1 - R_A \bar{s},$$

using $R_A = R_B$ yields

$$\pi_{1B} - \pi_{1A} = \frac{1-\alpha}{\alpha}(l_{1B} - 1),$$

which is nonpositive if and only if $l_{1B} \leq 1$, or

$$\frac{a(M)}{a(0)} \geq \Upsilon > 1 \tag{2.29}$$

by (2.28). By symmetry, for a type 2 firm, the profit differential $\pi_{2A} - \pi_{2B}$ is nonpositive if and only if (2.29) holds.

Therefore, we may conclude that, when Marshallian externalities are sufficiently large in comparison with trade costs that (2.29) holds, the dispersed configuration involving the agglomeration of M firms of each type within the same region is an equilibrium. Because of these externalities, the agglomeration of firms belonging to the same industry gives rise to *endogenous spatial inhomogeneities*, which allow for the existence of an equilibrium once such inhomogeneities are sufficiently strong. They generate new forces that are able to overcome the locational instability caused by the competitive price mechanism. This agrees with what we will see in Chapter 3.

2.3.3 Corollaries and Extensions

The spatial impossibility theorem per se does not preclude the agglomeration of all agents into a single region. However, we will see that this is a very unlikely outcome. Indeed, if a competitive equilibrium exists, then the theorem implies that there is no costly trade between regions. Hence, in the right-hand side of (2.11), the first three terms must be zero, so that

$$I = (R_A + R_B)(\phi_A + \phi_B),$$

thus implying that, if either $\phi_A > 0$ or $\phi_B > 0$, it must be that $R_A = R_B = 0$.

Corollary 2.1 *Assume that a competitive equilibrium exists in a spatial economy with a homogeneous space. If there is vacant land in one region, then the land rent must be zero in both regions.*

When all agents locate in region A, then region B is empty. Using Corollary 2.1, the equilibrium land rent must also be zero in region A. This is so only when no agent in the economy has a positive marginal utility or productivity for land – a situation that is very unlikely in practice.

In fact, Corollary 2.1 is a special case of a more general result that has farfetched implications. Summing (2.9) and (2.10) across firms and households in region A and using (2.5)–(2.7), we obtain

$$\begin{aligned} I(A, B) &= (\boldsymbol{p}_B - \boldsymbol{p}_A) \cdot (\boldsymbol{E}_{AB} - \boldsymbol{E}_{BA} - \boldsymbol{y}_{tA}) + (R_B - R_A)(s_{tA} + \phi_A - S) \\ &= (R_B - R_A)(\phi_A - S). \end{aligned}$$

Likewise, we get

$$\begin{aligned} I(B, A) &= (\boldsymbol{p}_A - \boldsymbol{p}_B) \cdot (\boldsymbol{E}_{BA} - \boldsymbol{E}_{AB} - \boldsymbol{y}_{tB}) + (R_A - R_B)(s_{tB} + \phi_B - S) \\ &= (R_A - R_B)(\phi_B - S). \end{aligned}$$

At a competitive equilibrium, neither $I(A, B)$ nor $I(B, A)$ can be positive. As a consequence, if there is no vacant land in the economy ($\phi_A = \phi_B = 0$), it follows that

$$I(A, B) = (R_A - R_B)S \leq 0 \quad \text{and} \quad I(B, A) = (R_B - R_A)S \leq 0,$$

thus implying that $R_A = R_B$. Alternatively, if there is some vacant land in, say, B, ($\phi_B > 0$), then $R_B = 0$ so that $I(A, B) = R_A(S - \phi_A) \leq 0$. This in turn implies that $R_A = 0$; hence, $R_A = R_B$.

Corollary 2.2 *If a competitive equilibrium exists in a spatial economy with a homogeneous space, then the land rent must be the same in all regions.*

This corollary has the following fundamental implication for us: *In a homogeneous space, the competitive price mechanism is unable to explain why the land rent is higher in an economic agglomeration* (such as a city, a central business district, or an industrial cluster) *than in the surrounding area.* This implies that no commuting, a special form of trade, may arise in this model.

Some further remarks are in order. First, using the same approach as Starrett (1978), we can extend the theorem (as well as the two corollaries) to any finite number of regions at the expense of heavy notation. Second, we have assumed for notational simplicity that each firm locates in a single region. The argument could be generalized to permit firms to run two distinct plants, one plant per region. The aggregate incentives to move, I, would now be defined in relation to the activities performed by plants and households in each region because each plant amounts to a separate firm (Koopmans 1957). Third, we have considered a closed economy. The model and the theorem can readily be extended to allow for trade with the rest of the world provided that each region has the same access to the world markets in order to satisfy the assumption of a homogeneous space. Last, the size of the economy is immaterial for the spatial impossibility theorem to hold, for assuming a "large economy," in which competitive equilibria often emerge as the outcome generated by several institutional mechanisms, does not affect the result because the value of total transport costs within the economy rises when agents are replicated.

Our main critical assumption regards the homogeneity of space. Relaxing this assumption may help to restore a competitive equilibrium involving transportation. To see it, consider the example of Figure 2.1 in which we introduce inhomogeneities in space. More precisely, we assume that location B has an exogenous attribute that makes the firm less productive there. This exogenous attribute may correspond to the absence of some immobile inputs used by the firm or to differences in land quality. In any event, if the firm locates in B, the set of feasible allocations is given by the new triangle $OE'F'$, as depicted in Figure 2.2. We see that the set of feasible allocations is now convex.

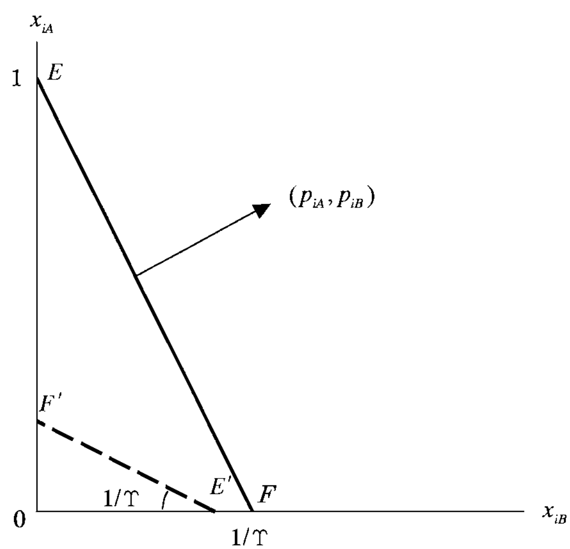


Figure 2.2: The set of feasible trade patterns in a heterogeneous space.

To gain more insights about the effect of spatial inhomogeneities, consider one more time our two-firm, two-location example with fixed lot size. We assume that location A has an exogenous attribute beneficial to firm 1 only (whereas firm 2 experiences a similar advantage in the sole location B).[12] Whatever the reason for it, this attribute gives rise to an additional earning equal to b_A when firm 1 is at A. Then, firm 1's profit at location A is now defined as follows:

$$\pi_{1A} = a_1 + b_A + p_{1A}q_1 - p_{2A}q_2 - R_A,$$

whereas π_{1B} is unchanged. Measuring again the incentive to move by the difference in profit at A and B, we obtain

$$\pi_{1B} - \pi_{1A} = t_1q_1 + t_2q_2 + R_A - R_B - b_A,$$

which is negative when b_A is sufficiently large. Because the same argument holds for firm 2, we may conclude that *a competitive equilibrium involving trade may well exist if firms have strongly diverging preferences for location attributes.* Or, to put it differently, the market breaks down when interagent transportation costs outweigh other aspects of agents' preferences about locations.[13] As Hamilton (1980, 38) put it,

> Stability is lent to the system by having plants differ from one another in their preferences for the sites *qua* sites, and instability arises from a large volume of trade among plants.

The spatial impossibility theorem is important because the competitive price mechanism is the keystone of the economic theory of market equilibrium. For our purpose, it seems that little can be learned about the formation of the economic landscape by appealing only to the competitive framework unless we consider an heterogeneous space. In this case, as will be shown in Chapter 3, the competitive model keeps its relevance. On the other hand, if we want to explain the emergence of economic agglomerations in an otherwise homogeneous space, we must explicitly consider a noncompetitive setting. Examples include (1) *market failures* such as (technological) externalities (see Chapters 6 and 8) or public goods (see Chapter 5) and (2) *imperfect competition* (see Chapters 7 and 9). Whatever the solution retained, if we want to explain how economic agglomerations are formed, we have to appeal to nature (the unevenness in the distribution of natural resources), to nonmarket institutions (externalities), or to an imperfectly competitive paradigm.[14]

2.3.4 Notes on the Literature

One of the first spatial competitive models is the *spatial price equilibrium* model formalized by Cournot (1838, chap. X). The sellers and buyers of a commodity are located at nodes of a transportation network. The issue is then to determine simultaneously the quantities supplied and demanded at each node and the local prices at which the commodity is supplied by the sellers and bought by the customers. The equilibrium is reached when the demand price equals the supply price plus the transport cost for all positive flows; if the demand price is less than the supply price plus the transport cost, then no trade flow occurs.

This problem has been revisited by Enke (1951), who found a formal connection with the theory of electric circuits and proposed a solution method based on this analogy. Soon after, Samuelson (1952) showed that the market equilibrium can be obtained as the solution to a mathematical program containing the celebrated Hitchcock–Koopmans transportation problem, whereas Beckmann (1952) provided a continuous version of these problems. This cross-fertilization between geographical economics and mathematical programming has generated many extensions dealing with different aspects of production and demand, which are summarized in Takayama and Judge (1971) and Mougeot (1978). Later on, Florian and Los (1982) demonstrated that spatial equilibrium models may also be solved by means of variational inequalities, thus opening the door to a new flow of contributions that are reviewed and extended in Nagurney (1993). A shorter synthesis may be found in Labys and Yang (1997). The main limit of all these models is that they do not allow for the locational choice of economic agents: both firms and households are supposed to be located.

In this perspective, the quadratic assignment model introduced by Koopmans and Beckmann (1957) may be viewed as the first serious attempt made to capture the locational choices of agents when commodities are traded between them. The difficulties encountered by these authors have triggered a very limited number of articles. Heffley (1972, 1976) has provided some useful complements to the work of Koopmans and Beckmann. A nice overview of the literature centered on the spatial price equilibrium and the quadratic assignment problem can be found in Schweizer (1986).[15]

Schweizer, Varaiya, and Hartwick (1976) have proposed a general equilibrium location model in which goods are to be traded at some given marketplaces. Berliant and Konishi (2000) have shown the existence of a competitive equilibrium involving trade and agglomeration when the exchange of commodities must occur in such trading centers. In addition, their model allows the number and locations of such centers to be endogenous through a noncompetitive mechanism. Opening a trading center involves a positive setup cost, and thus increasing returns are present in their conception. In equilibrium, the open marketplaces are the solution to the trade-off between setup costs and transport costs, which is discussed in Section 2.5 (see also Chapter 4).

A more general model has recently been developed by Berliant and Zenou (2000). They considered a two-stage, general equilibrium framework. Their model entails a continuum of heterogeneous workers and a finite number of heterogeneous firms. Each firm uses an area of land as well as one type of labor. Firms first choose their locations strategically, whereas prices are determined competitively. Consumers/workers then determine their residence, working place, and consumption bundle, taking the firm locations and the prices of goods as given. Showing the existence and (local) uniqueness of an equilibrium turns out to be a hard task in such a setting. Berliant and Zenou have combined bid rent techniques and the tools of modern general equilibrium theory to show existence and local uniqueness. This suggests the use of alternative new techniques to get existence of noncooperative/competitive equilibria in spatial models.

2.4 THE FIRST WELFARE THEOREM IN A SPATIAL ECONOMY

It has been claimed sometimes that the first welfare theorem does not necessarily hold in a spatial economy when firms exchange intermediate goods. If it did, this would make the case of the market really bad for the analysis of a spatial economy. We want to show here that this criticism is unfounded because it disregards the assumption that markets are complete in the standard general equilibrium model.

To illustrate our point, we consider again our two-firm, two-location example in which location A has an exogenous attribute beneficial to firm 1 only, whereas firm 2 experiences a similar advantage in location B only. In this case, we know that a competitive equilibrium may exist. But is it efficient? That is, does the location pair (A, B) maximize the social surplus given by

$$\mathcal{S} = a_1\delta_{1i} + a_2\delta_{2j} - t_1 q_1 - t_2 q_2,$$

where $\delta_{1i} = 1$ if firm 1 is located at A $(i = A)$ and zero at B; the definition of δ_{2j} is similar. We know that the competitive equilibrium, when it exists, involves firm 1 at location A and firm 2 at location B. Hence, the total surplus is maximized.

More generally, considering the two-region economy described in Section 2.3, we say that an allocation is *efficient* if no feasible allocation exists such that every consumer is at least as well off and at least one consumer is strictly better off.

Proposition 2.3 *Assume a two-region economy. If the preferences U_{hr} satisfy local nonsatiation for all consumers at each location, then any competitive equilibrium is efficient.*

Proof Assume that the competitive allocation is not efficient and denote a dominating allocation by $\hat{N}_A$, $\hat{N}_B$, $\hat{M}_A$, $\hat{M}_B$, $\hat{\boldsymbol{x}}_{hr}$ $(r = A, B)$, $\hat{\boldsymbol{y}}_{fr}$ $(r = A, B)$, $\hat{\boldsymbol{E}}_{AB}$, $\hat{\boldsymbol{E}}_{BA}$, $\hat{\boldsymbol{y}}_{tr}$ $(r = A, B)$, and $\hat{s}_{tr}$ $(r = A, B)$.

Consider the consumers in N_A. Those who stay in region A ($h \in N_A \cap \hat{N}_A$) at the dominating allocation face the same prices as before so that

$$\begin{aligned} &\boldsymbol{p}_A \cdot \hat{\boldsymbol{x}}_{hA} + R_A \hat{s}_{hA} - \boldsymbol{p}_A \cdot \boldsymbol{\omega}_h \\ &\quad \geq \sum_{r \in \{A,B\}} R_r \tilde{s}_{hr} + \sum_{r \in \{A,B\}} \sum_{f \in M_r} \theta_{hf} \pi_{fr} + \theta_{ht} \pi_t \end{aligned} \tag{2.30}$$

because consumers are at least as well off and local nonsatiation implies that the budget constraints are all binding at the competitive equilibrium. Similarly, for the consumers who are now established in region B ($h \in M_A - \hat{M}_A$), it follows that

$$\begin{aligned} &\boldsymbol{p}_B \cdot \hat{\boldsymbol{x}}_{hB} + R_B \hat{s}_{hB} - \boldsymbol{p}_B \cdot \boldsymbol{\omega}_h \\ &\quad \geq \sum_{r \in \{A,B\}} R_r \tilde{s}_{hr} + \sum_{r \in \{A,B\}} \sum_{f \in M_r} \theta_{hf} \pi_{fr} + \theta_{ht} \pi_t, \end{aligned} \tag{2.31}$$

for otherwise local nonsatiation would imply that each of these consumers would have chosen to reside in region B instead of A at the competitive equilibrium. The same conditions hold for the consumers in N_B.

Because at least one consumer is strictly better off, one inequality in (2.30) and (2.31) must be strict for at least one region; otherwise, this consumer would not maximize her utility. Summing across consumers and regions, we obtain

$$\begin{aligned} &\sum_{r \in \{A,B\}} \left(\boldsymbol{p}_r \cdot \sum_{h \in \hat{N}_r} (\hat{\boldsymbol{x}}_{hr} - \boldsymbol{\omega}_h) + R_r \hat{s}_{hr} \right) \\ &\quad > \sum_{r \in \{A,B\}} R_r S + \sum_{r \in \{A,B\}} \sum_{f \in M_r} \pi_{fr} + \pi_t. \end{aligned} \tag{2.32}$$

Because the dominating allocation is feasible, we may use the material balance conditions (2.5)–(2.7) evaluated at this allocation together with (2.32) to show that

$$\sum_{r \in \{A,B\}} \sum_{f \in M_r} \hat{\pi}_{fr} + \hat{\pi}_t > \sum_{r \in \{A,B\}} \sum_{f \in M_r} \pi_{fr} + \pi_t,$$

which means that the aggregate profits made by the firms and the carriers would be higher at the dominating feasible allocation. Consequently, some firms, the carrier, or both do not maximize profits at the competitive equilibrium, which is a contradiction. Q.E.D.

Note that Proposition 2.3 remains valid for any finite set of locations and can be extended to the case in which some agents undertake their activities in several regions.[16] As such, this result should not come as a surprise because the first welfare theorem does not require any convexity assumption and is, therefore, not affected by the assumption that consumers and producers establish themselves into a single (or a small number of) location(s).

Nevertheless, the first welfare theorem requires complete markets. This seems to be a very demanding assumption in a spatial setting. Indeed, it means that firms and households make all their decisions on the basis of publicly posted prices not only for the active markets but also for all the potential markets. Because a good or service is differentiated by its location, this means that firms and households know $2n$ prices, one price being associated with each good in each location (for simplicity land and transport prices are neglected). Stated differently, in order to figure out whether or not a consumer wants to buy (to sell) a good in her location, the consumer (the producer) must know the price of this good even at a location at which nobody chooses to buy (to sell) it in equilibrium. It seems hard to believe that something like that could happen because the price of a good is not quoted before a market is open. Yet, we know from Samuelson (1952) that a relationship exists between the equilibrium prices of the same physical good at different locations that allows one to obviate this difficulty. The equilibrium price at a location in which the good is not produced is equal to the minimum of the marginal production cost (at each place the good is produced) plus the corresponding transport cost. Hence, once firms and consumers know the matrix of shipping costs between any pair of locations, it is sufficient to know the prices where the good is actually produced to infer the equilibrium prices at all the other places.

2.5 CONSIDERATIONS ON THE SECOND WELFARE THEOREM IN A SPATIAL ECONOMY

In the quadratic assignment problem, we have seen that, in general, the optimum cannot be decentralized by a system of prices and land rents. In other words, the second welfare theorem does not hold in a spatial economy when there are indivisibilities in firms. However, this does not preclude the possibility of decentralization through a nonlinear price system involving both a positive fee and marginal cost pricing. A well-developed body of literature is available in operations research, called *facility location analysis*, which is of great interest to economists and geographers.[17] Operations researchers and management scientists have, indeed, extensively studied some basic location models, thus making it possible to gain more insight into the nature and magnitude of the difficulties posed by location problems involving indivisibilities.

2.5.1 The Simple Plant Location Problem

In facility location analysis, a particular model has emerged as the main prototype for situating facilities, such as industrial plants or warehouses, to minimize the cost of satisfying the demand for some commodity. In general, there are fixed costs for locating the facilities and transport costs for distributing the commodity from the facilities to the consumers. When there is no a priori capacity constraint on the facilities, this problem is commonly referred as the *simple plant location problem* (SPLP). This model encapsulates the fundamental trade-off between fixed production costs and transportation costs. As will be seen in Chapter 4, this trade-off is at the heart of many location models and is crucial for the geography of economic activities regardless of the particular institutional setting in which those activities develop.

Having said that, we may turn to the SPLP, which we now define: Given a spatial distribution of requirements for a (composite) commodity, the purpose of the model is to determine the number and locations of facilities so as to minimize the sum of production and transportation costs. On the demand side, social needs are expressed by some fixed requirements distributed over a finite number of given locations $j = 1, \ldots, N$; the requirement in site j is denoted by δ_j. On the supply side, facilities can be placed at a finite number of potential locations $i = 1, \ldots, M$, where production displays increasing returns. The setup cost F_i and marginal production cost c_i are constant; hence, the total production cost of a facility at i producing q_i units of the commodity is given by $F_i + c_i q_i$. Different fixed costs (F_i) may account for differences in fixed factors endowments, whereas the marginal costs (c_i) may reflect other specificities in local conditions of production. Finally, the cost of shipping one unit of the commodity from site i to site j is a constant t_{ij}. The matrix of transportation costs (t_{ij}) is general enough to account for various access conditions to the local markets j. Unlike the quadratic assignment problem, there is no direct interaction between plants in the SPLP. However, *facilities' locations are interdependent through the market areas they supply*. This model, therefore, allows for indirect interaction among facilities.

Formally, the SPLP is defined by the following mathematical program:

$$\min Z = \sum_{i=1}^{M} \sum_{j=1}^{N} (c_i + t_{ij}) \delta_j x_{ij} + \sum_{i=1}^{M} F_i y_i \tag{2.33}$$

subject to the constraints

$$0 \leq x_{ij} \leq y_i \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N \tag{2.34}$$

$$\sum_{i=1}^{M} x_{ij} = 1 \qquad j = 1, \ldots, N \tag{2.35}$$

$$y_i \in \{0, 1\} \qquad i = 1, \ldots, M, \tag{2.36}$$

where x_{ij} stands for the (nonnegative) fraction of the demand at j supplied by a facility at i, and y_i is a $0 - 1$ variable that equals 1 when a facility is set up at j and 0 otherwise. Clearly, the objective function (2.33) accounts for all the production and transportation costs to be borne in order to satisfy all the local requirements. The first set of constraints (2.34) implies that no demand can be supplied from a site where no facility has been built. The second set of constraints (2.35) means that the total requirement in each i must be met from the open facilities. Finally, the integral constraints (2.36) correspond to indivisibilities in facilities.

The SPLP is a linear program involving both continuous variables (x_{ij}) and integer variables (y_i). It can readily be verified that there is always an optimal solution such that any local market is supplied by a single facility, which implies that the constraints (2.34) can be replaced by the following integral constraints:

$$x_{ij} \in \{0, 1\} \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N,$$

and thus the SPLP becomes a linear program involving integer variables only.

We follow quite a different path in this section by assuming that the integral constraints (2.36) are replaced by the nonnegativity constraints

$$y_i \geq 0 \quad i = 1, \ldots, M, \tag{2.37}$$

where y_i will never exceed 1 at the optimal solution because $x_{ij} \leq 1$. In words, (2.37) means that we allow for a "fraction" y_i of a facility to be built in i. Stated differently, facilities are now assumed to be perfectly divisible in that the cost of setting a fractional facility of size y_i at i is just equal to $F_i y_i$; hence, there are constant returns to scale at the plant level. Given the constraints (2.34), at the optimum it must be that

$$y_i = \max_{j=1,\ldots,N} \{x_{ij}\} \quad i = 1, \ldots, M.$$

In this case, the SPLP becomes a standard linear program called the linear programming relaxation of the SPLP. Because the set of possible solutions is broader under (2.37) than under (2.36), the optimal value of the original problem is larger than or equal to that of the relaxed problem.

The dual of the linear programming relaxation is defined as follows:

$$\max Z_D = \sum_{j=1}^{N} \lambda_j \tag{2.38}$$

subject to the constraints

$$\lambda_j - \mu_{ji} \leq (c_i + t_{ij})\delta_j \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N \tag{2.39}$$

$$\sum_{j=1}^{N} \mu_{ji} \leq F_i \quad i = 1, \ldots, M \tag{2.40}$$

$$\mu_{ji} \geq 0 \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N, \tag{2.41}$$

where the λ_j are the dual variables associated with the constraints (2.35), whereas dual variables μ_{ji} are associated with the constraints (2.34) rewritten as follows:

$$y_i - x_{ij} \geq 0 \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N.$$

The economic interpretation of the dual variables is straightforward: λ_j is the (shadow) willingness to pay of the consumers located at j for having their requirement δ_j at j, whereas μ_{ji} is the (shadow) fee that the same consumers are willing to propose for a facility at location i to be set up. The set of consumers supplied by facility i may be interpreted as a club whose membership fee for consumers at j is precisely given by μ_{ji}. Consequently, maximizing (2.38) implies the maximization of the total revenue collected from all consumers, subject to the following constraints:

1. The offer made by consumers at location j, net of the fee they pay to get the commodity from i, never exceeds the cost of supplying those consumers from location i (see (2.39));
2. The sum of fees across consumers associated with facility i does not exceed the fixed cost of setting a facility at that location (see (2.40));
3. The fees μ_{ji} are nonnegative (see (2.41)), but there is no constraint on the offers λ_j.

With this interpretation in mind, it is easy to see that each facility may be interpreted as a separate agent that aims to maximize its profit defined by the following expression:

$$\pi_i = \sum_{j=1}^{N}[\lambda_j - (c_i + t_{ij})\delta_j]x_{ij} - y_i F_i \quad i = 1, \ldots, M,$$

where each bracketed term stands for the gross profits earned from selling in market j and the last term for the endogenous "fixed" cost the agent i must incur for operating.

It is well known from linear programming that the following complementary conditions hold at the optimum:

$$x_{ij}^* > 0 \Rightarrow \lambda_j^* - \mu_{ji}^* = (c_i + t_{ij})\delta_j \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N \tag{2.42}$$

$$y_i^* > 0 \Rightarrow \sum_{j=1}^{N} \mu_{ji}^* = F_i \quad i = 1, \ldots, M \tag{2.43}$$

$$\lambda_j^* > 0 \Rightarrow \sum_{i=1}^{M} x_{ij}^* = 1 \quad j = 1, \ldots, N \tag{2.44}$$

$$\mu_{ji}^* > 0 \Rightarrow y_i^* = x_{ij}^* \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N. \tag{2.45}$$

Condition (2.42) means that a positive flow from i to j signifies that the offer made by consumers at i, net of their contribution to the building of a facility at this location, is just equal to the actual cost of supplying these consumers with their requirement δ_j from a facility at i.[18] Condition (2.43) means that, at location i accommodating a facility, the sum of fees across consumers must exactly cover the fixed cost of opening a facility at i. When the offer made by consumers at i is positive, then (2.44) means that the commodity requirement at i is satisfied by the facility system. Finally, (2.45) means that a positive fee from consumers at j for a facility at i implies that the fraction of the requirement at j satisfied from i is equal to the "fraction" of facility open at i.

These conditions imply that, at the optimum, open facilities earn zero profits ($\pi_i = 0$), whereas the other facilities are not open because they would make negative profits ($\pi_i < 0$) at the equilibrium offers and fees. In other words, the economy works as if the commodity price at market j (defined by $(\lambda_j^* - \mu_{ji}^*)/\delta_j$) were equal to the marginal cost $(c_i + t_{ij})$ when market j is supplied by the facility at i and lump-sum transfers (μ_{ji}^*) are available to match exactly the fixed production costs of the open facilities. It is well known from linear programming that such dual variables always exist, and thus the optimum may be decentralized by prices and lump-sum transfers.

Unfortunately, the counterexample presented below shows that the solution to the dual may yield a solution whose primal involves fractional values for the variables y_i, thus violating the integral constraints (2.36). To make calculations easier, we substitute (2.39) into (2.40) and set

$$[\lambda_j - (c_i + t_{ij})\delta_j]^+ = \max\{\lambda_j - (c_i + t_{ij})\delta_j, 0\}$$
$$i = 1, \ldots, M \quad \text{and} \quad j = 1, \ldots, N$$

so that the dual program may be rewritten under the following condensed form:

$$\max \hat{Z}_D = \sum_{j=1}^{N} \lambda_j$$

subject to

$$\sum_{j=1}^{N} [\lambda_j - (c_i + t_{ij})\delta_j]^+ \leq F_i \quad i = 1, \ldots, M.$$

In this new setting, $[\lambda_j - (c_i + t_{ij})\delta_j]$ describes the fee that consumers at j are willing to pay in order to have a facility at i supplying them at its marginal production and transportation cost.

Example. The local markets and potential locations coincide with the vertices of a triangle whose sides have a length equal to 1. The commodity requirements are the same across markets with $\delta_j = 1$. The production cost is the same across locations with $F_i = 2$ and $c_i = 0$. Finally, transportation costs are linear in distance, and the commodity is shipped in a clockwise manner (the admissible routes are from 1 to 2, from 2 to 3, and from 3 to 1) so that the matrix (t_{ij}) is as follows:

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix}.$$

Clearly, it is less costly to supply, say, the consumers at 1 first from site 1, then from site 3, and finally from site 2. The corresponding condensed dual is:

$$\max \hat{Z}_D = (\lambda_1 + \lambda_2 + \lambda_3)$$

subject to

$$[\lambda_1]^+ + [\lambda_2 - 1]^+ + [\lambda_3 - 2]^+ \leq 2$$
$$[\lambda_1 - 2]^+ + [\lambda_2]^+ + [\lambda_3 - 1]^+ \leq 2$$
$$[\lambda_1 - 1]^+ + [\lambda_2 - 2]^+ + [\lambda_3]^+ \leq 2.$$

By symmetry, it is straightforward to check that

$$\lambda_1^* = \lambda_2^* = \lambda_3^* = 3/2,$$

and thus the optimal value of the dual is equal to $9/2$. Hence, the complementarity conditions (2.42)–(2.45) are satisfied if and only if

$$y_1^* = x_{11}^* = x_{12}^*$$

so that facility 1 supplies markets 1 and 2,

$$y_2^* = x_{22}^* = x_{23}^*,$$

facility 2 supplies markets 2 and 3,

$$y_3^* = x_{31}^* = x_{33}^*,$$

and facility 3 supplies markets 1 and 3. Because $x_{1j}^* + x_{2j}^* + x_{3j}^* = 1$, it then follows that $y_i^* = 1/2$ for all i. The preceding three equalities therefore imply that

$$x_{11}^* = x_{12}^* = x_{22}^* = x_{23}^* = x_{31}^* = x_{33}^* = 1/2$$
$$x_{13}^* = x_{21}^* = x_{32}^* = 0.$$

As a consequence, the relaxed problem yields a solution in which "half" a facility is established at every location. Total production costs are equal to 3, and total transportation costs to $3/2$. This result obviously contradicts the integer constraints (2.36) because there is no optimal solution in which each market is supplied from a single facility (e.g., x_{11}^* and x_{31}^* are both positive). Therefore, it is sufficient to build the fraction of the facility corresponding to the maximum of the fractions of the requirements it serves (here $1/2$) to get the optimum of the relaxed problem.

Going back to the original problem (2.33)–(2.36), some simple calculations show that the optimal solution is given by

$$y_1^* = y_2^* = 1$$

$$x_{11}^* = x_{22}^* = x_{23}^* = 1,$$

thus implying that total production costs are equal to 4 and total transportation to 1. Because plants are indivisible, more is spent on fixed production costs without permitting reduced transport costs. This yields a value for the objective function (2.33) equal to 5, which is strictly larger than the optimum value of the relaxed problem (9/2). The difference $Z - \hat{Z}_D$ is called the *duality gap*. It is precisely the existence of this gap that makes it problematic, not to say impossible, to decentralize the optimal configuration even though we allow for more instruments, that is, prices and fees.

The situation is not as bad as it seems to be, however. First, in practical problems, the solution to the dual of the relaxed SPLP is often an integer, thus making decentralization through a system of nonlinear prices possible. Second, there are some interesting recent results showing how the SPLP and the dual are connected, which are now discussed. Assume that marginal production costs are equal across locations and that transportation costs t_{ij} satisfy the triangular inequality

$$t_{ij} \leq t_{ik} + t_{kj} \quad \text{for all } i,\ j, \text{ and } k$$

as well as the symmetry property

$$t_{ij} = t_{ji}.$$

The last two conditions hold under fairly general conditions such as transport costs given by a concave and increasing function of any metric (Huriot, Smith, and Thisse 1989).

In this case, Chudak (1998) has shown that the optimal value of Z never exceeds the optimal value of Z_D by more than 74%:

$$\max Z_D \leq \min Z \leq [(e + 2)/e] \max Z_D. \tag{2.46}$$

In other words, a configuration of facilities and a two-part tariff system always exist that allow one to recover at least the fraction $e/(e + 2) \approx 0.576$ of the total cost incurred. It is worth stressing that this bound holds regardless of the form of the metric used to measure distance over space and for any finite set of locations and markets. Furthermore, it is conjectured that the upper bound used in (2.46) is not the best possible; for the time being, it seems that this bound would be about 1.463, in which case the system operator would be certain to recoup at least 68% of the total cost (Guha and Khuller 1998).

In many practical problems, however, the duality gap is often much lower than the one given in (2.46). In particular, when all the locations and the markets are colinear, it has been shown that

$$\min Z = \max Z_D,$$

which means that any one-dimensional SPLP may be decentralized by means of a nonlinear price system. This is a very nice result especially relevant for transport systems developed along a major corridor. However, this result is also somewhat misleading because it might provide wrong insights about decentralization within more general location problems.

Last, there is no duality gap if and only if the core of the economy corresponding the SPLP is nonempty (Goemans and Skutellla 2000). This result has two major implications:

1. Being able to decentralize the optimal solution to the SPLP means that there is a pricing rule such that no group of consumers located at some markets j may be strictly better off by seceding from the others.
2. The nonlinear pricing rule considered in the foregoing discussion belongs to the core and is therefore accepted by the grand coalition of all the consumers.

Those results suggest that approximate nonlinear competitive equilibria may exist since the approximation obtained by relaxing the integral constraints in the SPLP is often fairly good. We will return to the decentralization of the optimum, using a different approach, in Sections 4.3 and 5.2.

2.5.2 Notes on the Literature

The SPLP has been proposed more or less simultaneously by various authors, including Kuehn and Hamburger (1963), Manne (1964), and Stollsteimer (1963). Since then, it has attracted much attention in operations research, and a great deal has been learned about this model and its various extensions. Extensive surveys of the literature related to the SPLP can be found in Francis and Mirchandani (1990) and Labbé, Peeters, and Thisse (1995).

2.6 CONCLUDING REMARKS

Economic theory has long been dominated by the competitive paradigm, and this may partly explain why space has been neglected by the economics profession. Indeed, we have seen that, when a competitive equilibrium exists in a homogeneous spatial economy, no transportation can occur; hence, regions do not specialize and agglomerations of firms and households cannot be formed and sustained. This is not what we observe in the real world.

Of course, assuming that space is heterogeneous as a result of an uneven geographical distribution of natural resources may help to explain a nonuniform distribution of economic activities over space. It is our contention, however, that the diversity of resources does not provide a fully satisfactory explanation for the existing spatial concentration and regional specialization. In saying this, we do not deny the usefulness of neoclassical trade theory based on constant returns to scale and comparative advantage. We mean that, by relying solely on the diversity of resources over space, one puts aside the main endogenous socioeconomic forces that yield agglomeration and specialization ("second nature"). Even though physical attributes may explain why particular locations nest agglomerations such as cities, these attributes are not sufficient to explain why these agglomerations may become so large. As argued previously, their size must be explained through the interplay of economic and social interactions within economic models.[19] To perform such a task and to construct economically relevant theories of agglomeration, we must depart from the competitive paradigm.

Finally, because land is essential and scarce and transportation is costly in the real world, the presence of some kinds of indivisibilities is crucial to understanding the emergence of economic agglomerations (see also our discussion in Chapter 1). In this case, the decentralization of economic decisions by means of perfectly competitive markets is problematic – even when two-part tariffs are permitted. We will see, however, that much can be achieved by appealing to noncompetitive theories of market behavior and to nonmarket institutions.

APPENDIX

In the general case with M firms and M locations, when locations are appropriately situated (for example, locations are distributed along a circle) and the input–output linkages among firms are appropriately chosen, a location assignment may seem supportable by a competitive price system. And, indeed, since the publication of Koopmans and Beckmann (1957), several counterexamples have been proposed. After a close examination, however, they turn out to be false. To prevent the reader from making such a vain attempt, we show here that no competitive equilibrium exists in the quadratic assignment problem when space is homogeneous.

There are M firms ($i = 1, \ldots, M$) and M homogeneous locations ($r = 1, \ldots, M$). Firm i produces good i using inputs q_{ji} ($j = 1, \ldots, M$) produced by each other firm j, where the q_{ji}'s are constants such that at least one q_{ji} is strictly positive. By convention, we set $q_{ii} = 0$ for all i. Each location can accommodate one firm only. Let $t_i(r, s) > 0$ be the cost of shipping one unit of good i from r to s ($r \neq s$).

Consider any feasible assignment. Without loss of generality, we assume that firm i is assigned to location i. In this way, each location is characterized by the index of the firm that is assigned there. Let $\{p_i(r), R(r); i, r = 1, \ldots, M\}$ be the price system assumed to support this assignment. In this case, the profit of firm i at location i is given by

$$\pi_i(i) = a_i + p_i(i)\sum_{j=1}^{M} q_{ij} - \sum_{j=1}^{M} p_j(i)q_{ji} - R(i),$$

where a_i is a constant independent of the firm's location. If firm i relocates to location r and conducts the same activity, its profit is given by

$$\pi_i(r) = a_i + p_i(r)\sum_{j=1}^{M} q_{ij} - \sum_{j=1}^{M} p_j(r)q_{ji} - R(r),$$

Let us define the incentive for firm i to move from location i to location r by

$$I_i(i,r) = \pi_i(r) - \pi_i(i) \quad r = 1, \dots M \text{ and } r \neq i \tag{A.1}$$

and define the aggregate incentive for all firms to move (to all other possible locations) as follows:

$$I = \sum_{i=1}^{M}\sum_{r=1}^{M} I_i(i,r). \tag{A.2}$$

In equilibrium, because no $I_i(i,r)$ can be positive, the aggregate incentive I must be nonpositive.

In fact, we show below that

$$I = M\sum_{i=1}^{M}\sum_{j=1}^{M} q_{ij}t_i(i,j); \tag{A.3}$$

that is, the aggregate incentive for all firms to move is M times the total transport costs associated with the original assignment, which is strictly positive. This implies that no feasible assignment can be supported by a competitive price system.

To demonstrate (A.3), we proceed as follows. First, observe that in calculating $I_i(i,r)$ in (A.1), the constant a_i disappears. Second, in evaluating I as defined by (A.2), when we compute the sum $I_i(i,r) + I_r(r,i)$ for each pair (i,r), the land rents $R(i)$ and $R(r)$ disappear. Therefore, we may focus on quantities q_{ij} and express the total incentive I as follows:

$$I = \sum_{i=1}^{M}\sum_{j=1}^{M} q_{ij} f_{ij}(P), \tag{A.4}$$

where $f_{ij}(P)$ is a function of all prices $p_i(r)$. To determine these functions, let us focus on any one pair (i,j) such that $q_{ij} \neq 0$. Because q_{ij} is a part of the sales made by firm i while it is also an input of firm j, q_{ij} appears in each of the following incentive-to-move functions related to firm i and firm j:

$$\text{for firm } i\text{: } I_i(i, r) = \pi_i(r) - \pi_i(i) \qquad r \neq i,$$
$$\text{for firm } j\text{: } I_j(j, s) = \pi_j(s) - \pi_j(j) \quad s \neq j.$$

Therefore, in calculating the aggregate incentive I, the quantity q_{ij} appears only in the following subtotal:

$$\sum_{r=1}^{M} I_i(i, r) + \sum_{s=1}^{M} I_j(j, s) \equiv I_i(i, j) + I_j(j, i) + \sum_{\substack{r=1 \\ r \neq i, j}}^{M} \{I_i(i, r) + I_j(j, r)\}.$$

In $I_i(i, j)$, using (A.1) in which $r = j$, we see that the term related to q_{ij} appears as $q_{ij}[p_i(j) - p_i(i)]$. Similarly, in $I_j(j, i)$, the term related to q_{ij} appears as $q_{ij}[p_i(j) - p_i(i)]$; this stands for firm j's cost saving on input q_{ij} associated with firm j's relocation from j to i. Hence, in $I_i(i, j) + I_j(j, i)$, the term related to q_{ij} appears as $2q_{ij}[p_i(j) - p_i(i)]$.

Since, for each pair (i, j), the equilibrium prices must be such that

$$p_i(j) = p_i(i) + t_i(i, j) \quad \text{when } q_{ij} > 0$$

in $I_i(i, j) + I_j(j, i)$ we have

$$2q_{ij}[p_i(j) - p_i(i)] = 2q_{ij}t_i(i, j). \tag{A.5}$$

Finally, take any $r \neq i, j$. Then, in $I_i(i, r)$, the term related to q_{ij} appears as $q_{ij}[p_i(r) - p_i(i)]$, which represents the change in revenue for firm i from selling the quantity q_{ij} at location r instead of selling it at location i. On the other hand, in $I_j(j, r)$, the term related to q_{ij} appears as $q_{ij}[p_i(j) - p_i(r)]$, which represents a part of firm j's cost saving associated with the relocation from location j to r. Hence, in $I_i(i, r) + I_j(j, r)$, we have

$$\begin{aligned} q_{ij}[p_i(r) - p_i(i)] + q_{ij}[p_i(j) - p_i(r)] &= q_{ij}[p_i(j) - p_i(i)] \\ &= q_{ij}t_i(i, j), \end{aligned} \tag{A.6}$$

thus implying that, if both firms i and j relocate to $r \neq i$ and j, they can save the transport cost $q_{ij}t_i(i, j)$.

Because we have $M - 2$ such locations $r \neq i, j$, we see from (A.5) and (A.6) that, in the sum (A.4), the term related to q_{ij} appears as

$$\begin{aligned} 2q_{ij}[p_i(j) - p_i(i)] + (M - 2)q_{ij}[p_i(j) - p_i(i)] &= Mq_{ij}[p_i(j) - p_i(i)] \\ &= Mq_{ij}t_i(i, j), \end{aligned}$$

and thus, using (A.4), we obtain

$$f_{ij}(P) = M[p_i(j) - p_i(i)] = Mt_i(i, j).$$

Because this result holds for any pair (i, j) such that $i \neq j$, we may conclude as in (A.3).

NOTES

1. This idea was put forward by Allais (1943, 809).
2. This is consistent with a standard result in trade theory according to which perfectly competitive agents exchange commodities as long as there are differences in relative endowments, production techniques, or preferences across locations. Hence, no trade would occur in the noeclassical model with a homogeneous space (Samuelson 1939, 1962).
3. For example, if natural amenities may explain the development of the French Riviera, they do not seem to play any major role in the surge of a large metropolitan area of 30 million people such as Tokyo.
4. That convex preferences do not fit consumer behavior in a spatial economy had also been noticed by several general equilibrium theorists such as Hildenbrand (1974, 83) and Malinvaud (1970, 22).
5. More precisely, the proof of the existence of a competitive equilibrium assumes only the convexity of the production set of the economy, not the convexity of the production set of each firm, the latter being a sufficient condition for the former (Debreu 1959, chapter 6). This does not affect, however, the nature of the difficulty addressed here.
6. The resulting optimization problem involves a quadratic objective function, which accounts for the name of the problem.
7. The reader may find the competitive assumption unrealistic in the case of two firms. As shown by the spatial impossibility theorem proven below, the number of agents can be made very large without changing the conclusions.
8. For simplicity, we have not taken natural resources into account. However, they could be integrated into the model, if they are assumed to be uniformly distributed across regions but not (necessarily) across households.
9. Replacing equalities by inequalities in (2.5) and (2.6) does not affect the result, whereas the inequality in (2.7) turns out to be essential.
10. Of course, this expression underestimates the profit that firm f could make in region B by adjusting its production plan. The same holds for households. However, the argument used by Starrett is sufficient for the spatial impossibility theorem.
11. As shown by the vast literature developed around the theme of the "spatial price equilibrium," standard general equilibrium theory can be very useful for the study of commodity flows when both firms and households have fixed and given locations (see, e.g., Takayama and Judge, 1971).
12. Such a situation was typical of the manufacturing sector in the nineteenth century when plant locations were very much governed by the geographical distribution of raw materials, which was itself very uneven. This is much less relevant in modern economies replete with footloose firms.
13. See Heffley (1972, 1976) and Hamilton (1980) for further developments.
14. Ellickson and Zame (1994) disagreed with this claim and argued that the introduction of moving costs in a dynamic setting may be sufficient to save the competitive

paradigm. To our knowledge, however, the implications of their approach have not yet been fully worked out.

15. Fujita and Thisse (1993) have shown how the land capitalization process (see Chapter 4) may be used to provide a solution to a generalized version of the quadratic assignment problem. They designed a game in which firms' payoffs are defined by their profits augmented by the value of the land rent they create and showed that this game always has pure strategy equilibria that are also socially optimal.
16. However, in an economy with a continuum of locations and a continuum of consumers having location-dependent preferences, Berliant, Papageorgiou, and Wang (1990) have identified technical problems that cast doubt on the validity of the first welfare theorem in such a general context.
17. See Drezner (1995) for several surveys devoted to various aspects of this literature, whereas Hansen et al. (1987) have offered an economics-oriented survey.
18. These conditions are similar to those derived by Samuelson (1952) and discussed in Section 2.3.
19. For example, Cronon (1991) explained how first and second nature are to be combined to account for both the location and the growth of Chicago.

3

The Thünen Model and Land Rent Formation

3.1 INTRODUCTION

Land use models explain the way various activities using land locate over a given area. This phenomenon can be studied from a different perspective by asking which activities are accommodated in specific locations. As will be seen in this chapter, these two approaches may be considered interchangeable, although they differ somewhat. The first is more in line with microeconomics in that the analysis focuses on where given agents chose to locate, whereas the second is more akin to the approach followed by many geographers, who put the emphasis on places and densities and not on agents.[1]

Because, in a market economy, land is allocated among activities through the price of land, the land use problem is equivalent to asking how the price of land is determined in a competitive economy. This does not seem to be a feasible task, for we have just seen that the price mechanism does not work in a spatial economy. The spatial impossibility theorem does not preclude, however, the possibility of uncovering particular, but relevant, economic situations in which the price mechanism is able to govern the allocation of activities over space. This is precisely what we will try to do in this chapter.

The prototype of such particular situations has been put forward by Thünen (1826), who sought to explain the pattern of agricultural activities surrounding cities in preindustrial Germany. The various models developed in his footsteps can be cast within the Arrow–Debreu framework because transactions must occur at a given marketplace (the town in Thünen's analysis), whereas activities (the crops in Thünen's analysis) and land are supposed to be perfectly divisible.[2] Once markets are considered as perfectly competitive, it becomes easy to understand why the Thünian model has been extensively studied in both production theory and urban economics where it has proven to be a very powerful tool. That is, the Thünian model rests on the paradigmatic combination formed by the standard assumptions of constant returns and perfect competition, while assuming an exogenously located marketplace.

Each location in space is characterized by various factors such as soil conditions, relief, geographical position, and the like. Both land rent and land use vary across locations depending on these characteristics. Among them, the most important for location theorists is the transport-cost differential over space. Although Ricardo concentrated more on fertility differences in his explanation of the land rent, Thünen constructed a theory focusing on the transport-cost differentials across locations. For that, he used a very simple and elegant setting in which space is represented by a plain on which land is homogeneous in all respects except for a marketplace in which all transactions regarding final goods must occur. The location of this marketplace is supposed to be given, and the reasons for its existence are left outside of the analysis. (Possible explanations for the formation of such a marketplace will be dealt with in later chapters.) We will see that very interesting results can be obtained with this model. In essence, the Thünen model shows how the existence of a center is sufficient for a competitive land market to structure the use of space by different activities.

Not all transactions, however, need to occur at the market town. In particular, it seems reasonable to assume that intermediate inputs are traded on a local basis instead of being shipped to the marketplace. Therefore, we will extend the basic model by integrating intermediate goods, which are also produced from land but locally traded. This will allow us to shed light on the impact that technological linkages may have on the spatial distribution of activities.

Our purpose in this chapter is not to provide a comprehensive survey of what has been accomplished in the large body of land use theory. Instead, we have chosen to focus on the main principles underlying Thünen's analysis. To this end, we discuss in Section 3.2.1 the properties of a simple model formulated within the general competitive equilibrium framework. Specifically, we assume that (1) all agents are price-takers, (2) producers operate under constant returns to scale, and (3) there is free entry in each type of activity. The price-taking assumption in the land market can be justified on the grounds that land in a small neighborhood of any location belonging to a continuous space is highly substitutable, thus making the competitive process for land very fierce.

However, because our main concern is to determine which agent occupies a particular location, it appears to be convenient, both here and in subsequent chapters in which we work with a land market, to determine the land use equilibrium from the *bid rent function* suggested by Thünen. The concept of bid rent function is probably what makes Thünen's analysis of land use so original and powerful. In a sense, it rests on the idea that land at a particular location corresponds to a single commodity whose price cannot be obtained by the textbook interplay between a large number of sellers and buyers, for as Alonso (1964, 41) put it, "land as space is a homogeneous good and land at a location is a continuously differentiated good."

Having said that, our aim is to find what kind of spatial distribution may arise in equilibrium as well as the features of the land rent profile sustaining such a distribution. Though the model is very simple, it shows that the spatial heterogeneity generated by a preexisting center is sufficient to obviate the negative conclusion of the spatial impossibility theorem. Two extensions of the basic model will be considered, namely, the introduction of intermediate goods (Section 3.2.2) and the possibility of substitution between land and labor in production (Section 3.2.3).

In Section 3.3, we will continue our exploration of the Thünian model by studying its applications to the formation of the urban land rent and the residential distribution of housing within a monocentric city. In this case, as suggested by Isard (1956, chapter 8) and formally developed by Alonso (1964), the Thünian town is reinterpreted as the city center (or central business district) to which individuals must commute in order to work, whereas housing is developed in the surrounding area. Our main focus here will be on the households' trade-off between housing size and accessibility to the city center where jobs are available. We will see that this simple model provides a set of results consistent with the prominent feature of urban structures (Section 3.3.1). In particular, it explains the decrease in the urban land rent with distance away from the city center as well as the fall in the population density as one moves away from the center. As in the Thünian model, the city center plays a key role in the emergence of such a residential structure. Some comparative statics analysis is then performed on the residential equilibrium (Section 3.3.2). This analysis reveals several tendencies that agree with the main stylized facts suggested by urban economic history.[3] Among others, we note a spreading of the residential area corresponding to *suburbanization* when consumers get richer and commuting costs become lower, thus providing an explanation for what has been observed in many modern cities. We go on by showing that the market city is efficient in the absence of spatial externalities such as congestion in transport (Section 3.3.3).

In the foregoing analysis, the consumers are assumed to be identical in preferences and incomes. We go one step further by studying how the residential structure is affected when consumers are differentiated by their income (Section 3.3.4) and demonstrate that high-income consumers tend to settle far from the city center, which is left to the low-income ones. Finally, following the tradition of mainstream urban economics, we have assumed throughout this chapter a continuum of locations and consumers, thus working with a model in which *all* the unknowns are described by density functions. We show how the basic model of urban economics can be related to that of a city with a finite number of consumers located in a continuous urban space (Section 3.3.5). We conclude in Section 3.4 with a brief discussion of alternative but related urban models.

3.2 THE LOCATION OF DIVISIBLE ACTIVITIES

3.2.1 The Basic Model

By allocating an acre of land near the town to some crop, the costs of delivering all other crops are indirectly affected as they are forced to be grown farther away. Hence, determining which crops to grow where is not an easy task, thus making the work of Thünen very original. Though fairly abstract for the time, his treatment of the land use problem was not mathematical. One had to wait for the work of Launhardt ([1885], 1993, chap. 30), Lösch ([1940], 1954, chap. 5), and Dunn (1954) to have a formal presentation of his ideas.

The model is based on the following premises. There is a town located at the center of a featureless plain. All the products of various agricultural activities established in the surrounding area are to be traded there. The state formed by the town and its hinterland has no economic connections with the rest of the world; it is thus referred to as an *isolated state*. This isolated state is formally described by a large set of the Euclidean plane in which the town, treated as a point, is at the origin of the plane, whereas the distance from any point to the town is measured by the Euclidean distance. Each location r is identified by its distance r to the town.[4]

There are n activities, each producing a different agricultural good, or crop, denoted $i = 1, \ldots, n$. One may think of an activity as a set of farmers selling the same crop and using the same technology. The production of one unit of good i requires only the use of a_i units of land, where a_i is a positive constant independent of location, so that the technology of activity i exhibits constant returns to scale.[5] Consequently, if a unit of land at distance r is allocated to activity i, the corresponding production $q_i(r)$ of good i is given by

$$q_i(r) = \frac{1}{a_i}. \tag{3.1}$$

The density of land at each location is unity, and thus land density at distance r equals $2\pi r$.

Inasmuch as our focus is on land use, we put aside the determination of the prices of the agricultural goods in the town, which are supposed to be given and constant. Specifically, good i is sold at price p_i in the town to which it is shipped from its production place at a constant transport cost t_i per unit of good i and unit of distance. In other words, the product and transport markets are perfectly competitive.[6]

There is a perfectly competitive land market at every location in space, and the opportunity cost of land is assumed to be zero. However, as observed in the introduction, it is convenient to think that land at any point is allocated to an activity according to a bidding process in which the producer offering the highest bid secures the corresponding lot. In this regard, Thünen imagined a process in which each farmer makes an offer based on the surplus he can generate by using one unit of land available at any particular location. Because land is the only input and goods must be shipped to the market town, it should be clear that this surplus is given by $(p_i - t_i r)/a_i$. It varies with the activity but also with the location. Each activity i can then be characterized by a bid rent $\Psi_i(r)$ which is defined by the surplus per unit of land of any producer of good i at location r. Specifically, the *bid rent* of activity i at location r is here defined as follows:

$$\Psi_i(r) \equiv (p_i - t_i r)/a_i. \tag{3.2}$$

Since farmers are rational, they maximize profit per land unit. Being price-takers, the profit $\pi_i(r)$ made by a farmer in activity i per unit of land at location r is given by

$$\pi_i(r) = (p_i - t_i r)q_i(r) - R(r) = \Psi_i(r) - R(r),$$

using (3.1) and (3.2), where $R(r)$ is the rent per unit of land prevailing at distance r. Hence, if the profit earned by raising crop i at r is zero, the bid rent coincides with the market land rent.

In the present setting, a competitive equilibrium is defined by a land rent function and by the areas in which each activity is undertaken such that no producer finds it profitable to change the location of its activity at the prevailing land rent. Because returns to scale are constant, it follows that any farmer with a positive output earns zero profits, whereas the equilibrium land rent cannot be negative. Consequently, (3.2) implies that the equilibrium land rent is such that

$$R^*(r) \equiv \max\left\{\max_{i=1,\ldots,n} \Psi_i(r), 0\right\} = \max\left\{\max_{i=1,\ldots,n} (p_i - t_i r)/a_i, 0\right\} \tag{3.3}$$

so that the *land rent function* $R^*(\cdot)$ emerges as the upper envelope of the bid rent functions $\Psi_i(\cdot)$. In other words, at the end of the bidding, each location is occupied by the agent who is able to offer the highest bid.[7]

Each bid rent function being decreasing and linear in distance, we may conclude as follows:

Proposition 3.1 *The equilibrium land rent function is the upper envelope of all bid rent functions, and each crop is raised where its bid rent equals the equilibrium land rent. If the transport cost function is linear in distance, then the equilibrium land rent is decreasing, piecewise linear, and convex.*

As suggested in the introduction, it appears that the land rent is given by the *differential* surplus corresponding to the resources saved in transport by the most profitable activity relative to the zero surplus obtained at the extensive margin of land use. It even turns out that, for each activity, land rent is equal to the saving in

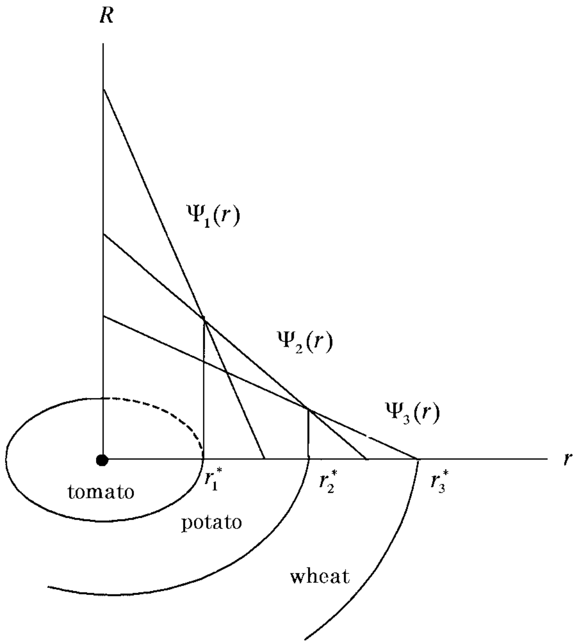


Figure 3.1: The land rent profile and Thünen rings when $n = 3$.

transport cost. This strict relationship should not be overemphasized, however, because it depends on the assumption of fixed technological coefficients (see Section 3.2.3). An illustration of a land rent profile in the case of three activities is provided in Figure 3.1.

It follows from Proposition 3.1 that, in equilibrium, the area allocated to a crop has the shape of a ring (or annulus) centered at the market town. Then, the crops, if they are raised, is ordered by the distance from the town in such a way that the crop having the steepest bid rent function locates nearest to the town, the crop with the second steepest locates in the next ring, and so on. Hence, it is not true that zones near the market town are necessarily locations of intensive type of land use or are appropriated by activities producing transport costly goods. Instead, as one moves away form the center, it is the activity with the steepest cost gradient, defined here by the ratio t_i/a_i, that outbids the remaining activities, and secures the corresponding location. For example, if the activities use more or less the same amount of land per unit of output, the hard-to-transport goods, typically because they are perishable, are produced close to the market town, whereas the easier-to-ship goods are produced farther away from their consumption place. Conversely, if the transport rates are about the same across goods, land-intensive activities are located close to the market town, whereas land-extensive activities are developed far from the center.

Proposition 3.1 implies several other things. First, all activities are distributed around the center which, therefore, appears to be the pivotal element in the spatial organization of production. Second, each location is specialized and activities are spatially segregated within rings of land. However, we will see that, in other contexts (see Section 3.2.2), integrated configurations in which different activities are undertaken at the same locations may also arise in equilibrium. Last, any activity k such that

$$\Psi_k(r) < R^*(r) \quad \text{for all } r \geq 0$$

has a zero output in equilibrium because it is unable to generate a surplus large enough to outbid the other activities anywhere in the plane. For notational simplicity, we assume from now on that all activities have a positive output in equilibrium.

Without loss of generality, we can reindex the activities in decreasing order of the slope (in absolute value) of their bid rent functions:

$$t_1/a_1 \geq \ldots \geq t_n/a_n.$$

We now show how the land use equilibrium pattern can be determined by using the bid rent function. Because activity 1 generates the highest surplus in the immediate vicinity of the town, it uses a disk of land (that is, an annulus with a zero inner radius) whose radius r_1^* must satisfy

$$\Psi_1(r_1^*) = \Psi_2(r_1^*);$$

that is,

$$r_1^* = \frac{p_1/a_1 - p_2/a_2}{t_1/a_1 - t_2/a_2}$$

beyond which activity 2 is undertaken because its surplus becomes higher than that of activity 1. Similarly, activity i $(= 2, \ldots, n-1)$ will occupy a ring whose inner radius r_{i-1}^* is such that

$$\Psi_{i-1}(r_{i-1}^*) = \Psi_i(r_{i-1}^*)$$

whereas the outer radius r_i^* is the unique solution to

$$\Psi_i(r_i^*) = \Psi_{i+1}(r_i^*)$$

because the two bid rents are to be equal along the border between two adjacent rings. Solving this equation yields

$$r_i^* = \frac{p_i/a_i - p_{i+1}/a_{i+1}}{t_i/a_i - t_{i+1}/a_{i+1}}.$$

Finally, the external margin of land use is endogenously determined at the distance r_n^* from the market town at which

$$\Psi_n(r_n^*) = 0$$

because the opportunity cost of land is assumed to be zero:

$$r_n^* = \frac{p_n}{t_n}$$

so that land is used only within a bounded disk whose radius is given by r_n^*. Beyond this distance stands Thünen's wilderness.

Since the equilibrium is competitive and there are no externalities, one expects this concentric pattern of rings to be socially optimal. That is, any other pattern in terms of size and shape would result in a lower *social surplus* S defined as the sum of crop values minus transport costs.

$$S \equiv \sum_{i=1}^{n} p_i Q_i - \sum_{i=1}^{n} T_i, \tag{3.4}$$

where Q_i is the output of activity i, and T_i is the corresponding transportation cost. Let $\theta_i(r) \geq 0$ denote the proportion of the land used by activity i at distance r ($\sum_i \theta_i(r) \leq 1$). Then, because $2\pi r$ units of land are available at distance r, we have

$$Q_i = \int_0^\infty \theta_i(r) 2\pi r / a_i dr$$

and

$$T_i = \int_0^\infty [\theta_i(r) 2\pi r / a_i] t_i r dr.$$

Substituting Q_i and T_i into (3.4) and using (3.2), we obtain

$$S = 2\pi \int_0^\infty \left[\sum_{i=1}^{n} \theta_i(r) \Psi_i(r) \right] r dr.$$

Maximizing S with respect to $\theta_i(\cdot)$ is therefore equivalent to maximizing the bracketed term at each location r with respect to $\theta_i(r)$ subject to $\sum \theta_i(r) \leq 1$. Clearly, activity i is carried out at distance r if and only if $\Psi_i(r)$ is positive and the maximum of all bid rents. Therefore, the optimum land use and market outcome are identical in the Thünian model, and both result in identical concentric annuli.[8]

The preceding analysis can be readily extended to the case of several production factors if production functions are of the fixed coefficient variety and if the return of each factor other than land is the same across locations. The case of a neoclassical technology is more complex and will be studied in Section 3.2.3.

The Thünian model can be closed by assuming that all agricultural activities need land and labor while a $(n + 1)$th manufactured good is produced in town by using labor alone – typically under the form of craftsmanship; such a specialization of tasks reflects the traditional division of labor between cities and the countryside. Workers are perfectly mobile and landlords reside in town; they all have identical (homothetic) preferences defined over the $(n + 1)$ goods. The solution to such a general spatial equilibrium model, in which the real wage common to all workers as well as the prices of agricultural and manufactured goods are endogenous, has been studied by Samuelson (1983) when $n = 2$ and by Nerlove and Sadka (1991) when $n = 1$. In Chapter 10, we follow a more general research strategy in which the formation of the town itself is made endogenous.

3.2.2 Technological Linkages and the Location of Activities

So far we have assumed that any produced good is shipped to the market town in which it is consumed. A well-known difficulty encountered in economics is to account for the existence of intermediate goods. It is interesting to figure out what the ring-shaped pattern obtained in the Thünian model becomes when some goods are used as inputs in the production of other goods. To the best of our knowledge, this problem has been first modeled by Mills (1970; 1972a, chap. 5) and extended further by Goldstein and Moses (1975). These authors assumed that intra-area shipments go by the shortest route and need not be shipped through the town.

The main change in the spatial organization of production is that several goods may be produced simultaneously at the same location instead of being produced in separated locations, as in the preceding section. To illustrate the working of such an economy, we adopt a slightly modified version of Mills by assuming that only two goods are involved, good 2 being used only as an input for producing good 1, which is itself shipped to the market town for being sold at a given price p_1. We will study this particular model in detail because it will allow us to see how all the equilibrium conditions interact to determine the equilibrium configuration and why the assumption of complete markets is needed.

As before, the production of one unit of good i requires a fixed amount of land a_i. However, producing one unit of good 1 requires also b units of good 2. Without loss of generality, the units may be chosen for $b = 1$.

If Q_1 is the quantity of good 1 produced and Q_2 the quantity of good 2 required, we have $Q_1 = Q_2$. It is worth noting that the equilibrium distance to the external land margin r_2 depends on the total production of good 1 but not on the way land is allocated between the two activities. Indeed, we have

$$a_1 Q_1 + a_2 Q_2 = (a_1 + a_2) Q_1 = \pi r_2^2$$

so that

$$r_2 = [(a_1 + a_2) Q_1 / \pi]^{1/2}.$$

An equilibrium configuration arises when no producer wants to change the location of his activity at the prevailing land rent and factor prices and when the spatial price equilibrium conditions for the intermediate product hold. Because the model is linear, we may disregard the intermediate cases and focus on the following two polar configurations: the integrated one, where both activities are undertaken together at each location, and the segregated one, where the two activities are separated as in Section 3.2.1. The spatial price equilibrium conditions then imply that it is never profitable to transport good 2 when the configuration is integrated; when the configuration is segregated, they say that the price of good 2 at any location where good 1 is produced is equal to the cost for one unit of good 2 to be available at the border between the two areas plus the transport cost from the border point to the production point.

To identify the conditions under which each configuration emerges as an equilibrium, it is again useful to work with the bid rent function associated with each activity. If $p_2^*(r)$ stands for the equilibrium price of good 2 at r, the surplus per unit of land (or the bid rent) of activity i at each r is defined as follows:

$$\Psi_1(r) = \frac{(p_1 - t_1 r) - p_2^*(r)}{a_1} \tag{3.5}$$

$$\Psi_2(r) = \frac{p_2^*(r)}{a_2}. \tag{3.6}$$

First, consider an integrated configuration. In this case, the two activities must have the same bid rent at each $r \leq r_2$, as illustrated in Figure 3.2. That is,

$$\frac{(p_1 - t_1 r) - p_2^*(r)}{a_1} = \frac{p_2^*(r)}{a_2}$$

or

$$p_2^*(r) = \frac{a_2}{a_1 + a_2}(p_1 - t_1 r). \tag{3.7}$$

Setting $\Psi_2(r) = 0$ (or $p_2^*(r) = 0$) at r_2, we obtain the fringe distance as follows:

$$r_2^* = p_1/t_1.$$

The integrated configuration is an equilibrium if and only if shipping good 2 must never be profitable. Because (3.7) is linear in distance, this amounts to

$$t_2 \geq \left| \frac{dp_2(r)}{dr} \right| = \frac{a_2 t_1}{a_1 + a_2}$$

or

$$\frac{t_2}{t_1} \geq \frac{a_2}{a_1 + a_2}. \tag{3.8}$$

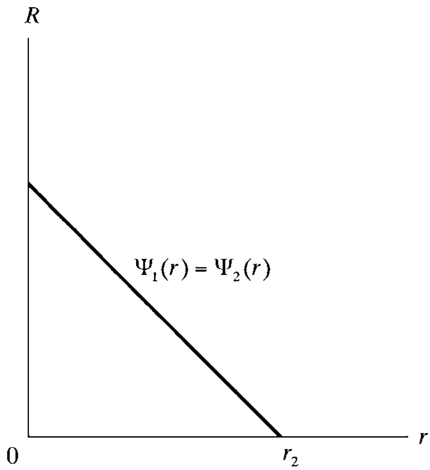


Figure 3.2: The rent profile for the integrated configuration.

This condition means that the cost of shipping one unit of good 2 is high relative to that of shipping one unit of good 1 given the relative intensity of land use in producing the two goods, and thus it is preferable to save on the transport of 2 than on the transport of 1.

The case of a segregated configuration is a bit more involved. Assume as in Figure 3.3 that good 1 is produced up to r_1, whereas good 2 is produced beyond r_1 up to r_2. Because the market for good 2 is competitive, everything works as if there were a marketplace for good 2 located in town, where this good is sold at some equilibrium price p_2^*. When good 2 is used at r, we have

$$p_2^*(r) = p_2^* - t_2 r. \tag{3.9}$$

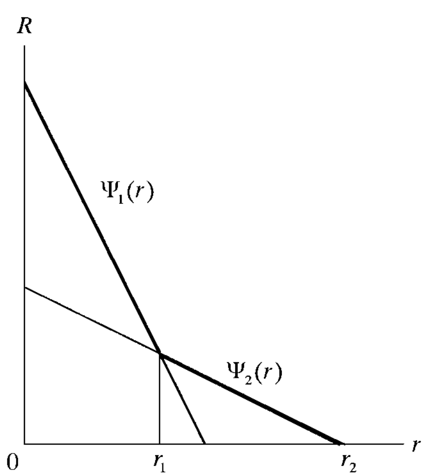


Figure 3.3: The rent profile for the segregated configuration.

Substituting (3.9) into (3.5) and (3.6) yields

$$\Psi_1(r) = \frac{(p_1 - p_2^*) - (t_1 - t_2)r}{a_1}$$

$$\Psi_2(r) = \frac{p_2^* - t_2 r}{a_2}.$$

The three unknowns, p_2^*, r_1, and r_2, can be determined by using the following equilibrium conditions. First, the two activities have the same bid rent at r_1:

$$\frac{(p_1 - p_2^*) - (t_1 - t_2)r_1}{a_1} = \frac{p_2^* - t_2 r_1}{a_2}.$$

Second, the bid rent of activity 2 is zero at r_2:

$$r_2 = p_2^*/t_2.$$

Third, $Q_1 = Q_2$ implies

$$\frac{\pi r_1^2}{a_1} = \frac{\pi\left(r_2^2 - r_1^2\right)}{a_2},$$

and thus

$$r_2 = \left(\frac{a_1 + a_2}{a_1}\right)^{1/2} r_1 \equiv k r_1,$$

where $k > 1$. The three conditions above yield

$$r_1^* = \frac{a_2 p_1}{(k-1)(a_1 + a_2)t_2 + a_2 t_1},$$

$$r_2^* = \frac{a_2 k p_1}{(k-1)(a_1 + a_2)t_2 + a_2 t_1},$$

and

$$p_2^* = k t_2 r_1^*.$$

For the segregated configuration to be an equilibrium, as shown in Figure 3.3, the bid rent curve of crop 1 must intersect that of crop 2 from above at distance r_1^*, thus implying

$$-\frac{d\Psi_1(r)}{dr} \geq -\frac{d\Psi_2(r)}{dr}.$$

This amounts to

$$\frac{t_1 - t_2}{a_1} \geq \frac{t_2}{a_1}$$

or

$$\frac{t_2}{t_1} \leq \frac{a_2}{a_1 + a_2}. \tag{3.10}$$

To sum up, we have shown:

Proposition 3.2

1. *If*

$$\frac{t_2}{t_1} \geq \frac{a_2}{a_1 + a_2}$$

holds, then the integrated configuration is an equilibrium.

2. *If*

$$\frac{t_2}{t_1} \leq \frac{a_2}{a_1 + a_2}$$

holds, then the segregated configuration is an equilibrium.

Though an equilibrium exists, it involves positive transport costs whatever its shape. This does not contradict the spatial impossibility theorem because the existence of a center turns out to be a major spatial inhomogeneity. Note also that the equilibrium may be characterized by positive interactivity transport costs. This is so when the cost t_2 of shipping the intermediate good to the producers of good 1 is low relative to the cost t_1 of shipping the final product to the market town. In this case, the equilibrium involves specializing land in the production of good 1 in the vicinity of the center, whereas good 2 is produced farther away; the pattern of production is ring-shaped as in the Thünian model. Otherwise, the two activities are spatially integrated to save the interactivity transport costs. Hence, in the presence of intermediate goods, the equilibrium does not necessarily involve spatial specialization.[9] In addition, there is no outward shipment of goods in equilibrium: either good 2 is consumed on the spot (as in the case of an integrated configuration) or transported toward the inner ring (as in the case of a segregated configuration).

Consequently, when there are technological linkages, the type of spatial configuration emerging at the market solution varies with the relative value of the transportation rates. This has an important implication: The fall in transport costs observed since the beginning of the Industrial Revolution does not imply that activities become indifferent with respect to their location. Even though transport costs would decrease, what matters for the organization of space is their relative changes.

The set of equilibrium patterns becomes richer once we allow for a more general input–output structure and relax the assumption of an isolated state by permitting imports through the market town at given prices p_1 and p_2 (Goldstein and Moses 1975). For example, when each activity uses the output of the other, if there are no imports, the inner ring is specialized in activity 1, whereas the outer ring involves integration: good 2 is produced for use in the first ring, but also for producing good 1 locally in the second ring, which, in turn, is used as a local input for producing good 2. When they compared their approach with the quadratic assignment model, Goldstein and Moses (1975, 77) were right when they claimed the following:

> By setting up a model with two goods, and a marketing center we are able to reach an equilibrium with complete interdependence and positive transport costs.

It is thus fair to say that the continuous approach to land use leads to important results with nontrivial equilibria. Unfortunately, the corresponding models become quickly intractable when the number of goods increases owing to the many possible special cases involved in characterizing equilibria.

The work by Koopmans and Beckmann (1957) has been at the origin of a long-standing debate about the (im)possibility of decentralizing the optimal configuration in a spatial economy. Of course, for this question to be addressed properly, one must work within a framework in which nontrivial competitive equilibria exist. In this perspective, Proposition 3.2 offers an interesting starting point. Furthermore, Mills (1970; 1972a, chap. 5) also showed that, in the model discussed above, the integrated solution is socially optimal if and only if (3.8) is verified, that is, when it pays to save on the transport of the intermediate good despite the need of shipping the final good to the center. On the contrary, when (3.10) holds, it is the segregated configuration that is socially desirable because it now pays to economize on the cost of shipping the final product. Accordingly, the optimum can be sustained as an equilibrium, and vice versa.

This turns out to be a fairly general property because Schweizer and Varaiya (1976) have been able to show that, in a monocentric economy, the optimal configuration can always be sustained by a decreasing and convex land rent in the general case of n goods with any input – output technology. This result is equivalent to the second welfare theorem for a spatial economy with divisible activities and technological linkages. Accordingly, we may safely conclude that the presence of intermediate goods does not prevent the existence of a competitive equilibrium when activities are perfectly divisible and when a single marketplace exists for some goods. In addition, the analysis of Mills reveals that any equilibrium is an optimum, that is, the first welfare theorem also holds (see also Goldstein and Moses 1975). Again this seems to be fairly robust in the case of divisible activities, though a general result comparable to Schweizer and Varaiya (1976) is missing (see, however, Proposition 2.1). This robustness, which is due to the divisibility of activities, makes the accessibility of an activity to the others potentially free because an integrated configuration is always feasible, whereas the existence of a single marketplace is a spatial heterogeneity that is used as a coordination device. In this case, there is no market failure. We will see an example of such a result in Section 3.3.3.

3.2.3 The Case of a Neoclassical Technology

Even though Thünen is considered the founder of marginalism, his model still belongs to the realm of classical economics to the extent that it assumes fixed technological coefficients. A more modern approach is obtained once substitution between land and labor, say, is allowed. This problem was tackled by Beckmann (1972a), who considered the case of a neoclassical Cobb–Douglas production function, but more general production functions could be similarly considered. Here we present a slightly more general analysis of this problem in that the parameter of this function may vary across activities. We assume that the assumptions of Section 3.2.1 are still valid, but (3.1) is now replaced by

$$q_i(r) = f[x_i(r)] = [x_i(r)]^{\alpha_i},$$

where $x_i(r)$ denotes the quantity (formally, the density) of labor units used per unit of land, whereas $q_i(r)$ is the output of good i per unit of land. In this expression, $0 < \alpha_i < 1$ stands for the substitution parameter between land and labor for good i. Hence, the marginal productivity of labor is positive and decreasing; the marginal productivity of land, given by $f(x_i) - x_i f'(x_i)$, is also positive and decreasing.

The profits $\pi_i(r)$ per unit of land earned by a producer at location r are then given by

$$\pi_i(r) = (p_i - t_i r)q_i - wx_i - R(r), \tag{3.11}$$

where w is the wage rate that is, for simplicity, supposed to be given and fixed, and constant across locations. Therefore, the corresponding profit-maximizing level of employment is

$$x_i^*(r) = \left[\frac{\alpha_i(p_i - t_i r)}{w}\right]^{\frac{1}{1-\alpha_i}} \quad \text{for } r \leq \frac{p_i}{t_i}. \tag{3.12}$$

Accordingly, for each activity, less and less labor is used as one moves away from the market town so that the equilibrium output is decreasing and continuous in the distance to the market town. Plugging (3.12) into (3.11) and setting $\pi_i(r) = 0$ and $R(r) = \Psi_i(r)$, we determine the maximum surplus that activity i may generate at location r. Consequently, the bid rent function associated with this activity is now defined by

$$\Psi_i(r) = (1 - \alpha_i)(\alpha_i/w)^{\beta_i}(p_i - t_i r)^{1+\beta_i} \quad \text{for } r \leq \frac{p_i}{t_i},$$

where $\beta_i \equiv \alpha_i/(1 - \alpha_i) > 0$. Hence, each bit rent function is decreasing and strictly convex in distance.

Without loss of generality, let $p_1/t_1 \leq \ldots \leq p_n/t_n$. Using the same argument as in Section 3.2.1, it may be shown that the equilibrium land rent is now given by

$$R^*(r) \equiv \max\left\{\max_{i=1,\dots,n}\Psi_i(r), 0\right\}$$
$$= \max\left\{\max_{i=1,\dots,n}(1-\alpha_i)(\alpha_i/w)^{\beta_i}(p_i - t_i r)^{1+\beta_i}, 0\right\}$$
$$\text{for } p_{j-1}/t_{j-1} \le r \le p_j/t_j,$$

and thus:

Proposition 3.3 *If production is described by a linear Cobb–Douglas function and if the wage rate is constant across locations, then the equilibrium land rent is decreasing and strictly convex in distance to the market town.*

Hence, using a neoclassical production function does not affect the general pattern of location, which is still described by a set of concentric rings,[10] whereas the land rent keeps the same decreasing and convex shape as in the Thünian model.

However, the simple and elegant condition describing the sequence of land use zones in the Thünian model does not hold any longer. The most surprising result, perhaps, is that the employment level may not be a continuous and decreasing function across activities. We have seen that this function is continuous and decreasing within each ring, but this does not necessarily hold at the border between two adjacent activities. Indeed, the equilibrium conditions imply that, at any distance r where activity i is undertaken, the land rent equals the marginal productivity of land whereas the wage equals the marginal productivity of labor, that is,

$$R(r) = (1-\alpha_i)[x_i^*(r)]^{\alpha_i}(p_i - t_i r)$$

as well as

$$w = \alpha_i[x_i^*(r)]^{\alpha_i - 1}(p_i - t_i r).$$

Taking the ratio of these two expressions yields

$$\frac{R(r)}{w} = \frac{x_i^*(r)}{\beta_i}.$$

Because, at the border r_i^* between the ith and $(i+1)$th rings, the same relationship holds for activity $i+1$, and since $R(r)/w$ is the same, it follows that

$$\frac{x_i^*(r_i^*)}{\beta_i} = \frac{x_{i+1}^*(r_i^*)}{\beta_{i+1}}.$$

Hence, the employment level is continuous across activities ($x_i^*(r_i^*) = x_{i+1}^*(r_i^*)$) if and only if the coefficients β_i are the same for all activities, that is, the production functions are identical for all activities. In this case, the equilibrium employment is a continuous and decreasing function of the distance to the market town across locations and activities.

On the other hand, if the coefficients α_i differ across activities, there is a discontinuity in the employment level at the border between two adjacent rings. Nevertheless, this input may still be decreasing. Let us check when this is so. For $x_i^*(r_i^*) > x_{i+1}^*(r_i^*)$ to hold, it follows that $\beta_i < \beta_{i+1}$, that is, $\alpha_i > \alpha_{i+1}$. Therefore, in equilibrium, the labor input is decreasing (but not continuous) provided that the locations of activities are ordered by decreasing order of the share of labor in the production of goods. There is no reason to expect this condition to be satisfied at the equilibrium configuration. Though the consumption of land remains specialized and ring-shaped, it therefore appears that the employment level may jump up or down when land use shifts from one activity to the next once substitution between land and labor is allowed.

Note, finally, that the inspection of the market land rent $R^*(r)$ reveals that, for any given activity, the decrease in the land value no longer fully compensates for the corresponding increase of the transport cost. The change in land price now induces a substitution from labor to land as one moves away from the market town, thus making this relationship more involved.

3.2.4 Notes on the Literature

The Thünian model has been formalized by Launhardt ([1885], 1993, chap. 30), Lösch ([1940], 1954, chap. 5), and Dunn (1954). Since then, much attention has been devoted to the possible reswitching of technologies as one moves away from the market town. The main results can be found in Schweizer and Varaiya (1976) and Schweizer (1978). A recent analysis of Thünen's original work is contained in Huriot (1994). Another domain of application of the Thünian model lies (somewhat ironically) in the neo-Ricardian models of production considered by Scott (1976) and Huriot (1981; 1994), among others.

3.3 THE URBAN LAND RENT

3.3.1 Residential Equilibrium in the Monocentric City

The basic urban model focuses on the fundamental trade-off between accessibility and space in residential choice, as developed by Alonso (1964), Mills (1967), and Muth (1969). To illustrate how it works, we consider a monocentric city with a prespecified center, called the central business district (CBD), where all jobs are located. For simplicity, the CBD is treated as a point, and space is assumed to be homogeneous except for the distance to the CBD. In this context, the only spatial characteristic of a location is its distance from the CBD, and thus the model is essentially one-dimensional. Compared with the Thünian model presented in Section 3.2.1, it therefore appears that the CBD replaces the market town, whereas the land available for raising crops is now used for housing.

Consider a continuum N of identical workers/consumers commuting directly to the CBD where they earn a given fixed income Y. Each consumer has a utility U depending on the quantity z of a composite good, which is available everywhere at a price equal to 1, and the lot size s of housing.[11] It is assumed that U is strictly increasing in each good, twice continuously differentiable, and strictly quasi-concave while both z and s are essential goods (every indifference curve has each axis as an asymptote). Furthermore, the lot size s is assumed to be a normal good. If a consumer is located at a distance r from the CBD, his budget constraint is then given by $z + R(r)s + T(r) = Y$, where $R(r)$ is the rent per unit of land at r and $T(r)$ the commuting costs at r. We suppose that there is no congestion in commuting while $T(r)$ is strictly increasing in distance and $0 \leq T(0) < Y < T(\infty)$.

The residential problem of the consumer can then be expressed as follows:

$$\max_{r,z,s} U(z, s), \quad \text{s.t.} \quad z + sR(r) = Y - T(r), \tag{3.13}$$

where $Y - T(r)$ is the net income at r. The only difference from the standard consumer problem is that here the consumer must also choose a residential location $r \geq 0$, which affects the land rent he pays, his commuting cost, and his consumption bundle. It should be clear that this problem encapsulates the trade-off between accessibility, measured by $T(r)$, and the land consumption, measured by s.

Because consumers are identical in terms of preferences and income, in equilibrium they must reach the same utility level u regardless of location. In the same spirit as in the Thünian model, we define the bid rent function $\Psi(r, u)$ of a consumer as the maximum rent per unit of land that he is willing to pay at distance r while enjoying a given utility level u. Given (3.13), we have

$$\begin{aligned} \Psi[Y - T(r), u] &\equiv \Psi(r, u) \\ &= \max_{z,s} \left\{ \frac{Y - T(r) - z}{s} \quad \text{s.t.} \quad U(z, s) = u \right\}. \end{aligned} \tag{3.14}$$

Indeed, for the consumer residing at distance r and selecting the consumption bundle (z, s), $Y - T(r) - z$ is the money available for land payment, and thus $[Y - T(r) - z]/s$ represents the rent per unit of land at r. The bid rent $\Psi(r, u)$ is then obtained when this rent is maximized by choosing the appropriate consumption bundle (z, s) subject to the constraint $U(z, s) = u$.[12]

Before continuing, two remarks are in order. First, observe the difference with the bid rent defined by (3.2) in the Thünian model in which it is implicitly assumed that the equilibrium profit level of activity i is zero. By contrast, the equilibrium utility level is endogenous here, making the land market across locations interdependent. Second, the bid rent function approach followed here is essentially the same as the indirect utility function approach used by Solow

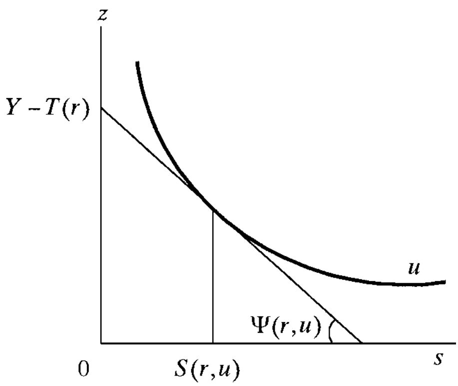


Figure 3.4: The equilibrium consumption bundle at r.

(1973) and is, therefore, closely related to duality theory as developed in microeconomics.

Because U is strictly increasing in z, we can always define the quantity $Z(s, u)$ of the composite good z, which solves $U(z, s) = u$. A standard analysis shows that $Z(s, u)$ is strictly decreasing, strictly convex in s such that $\lim_{s \to 0} Z(s, u) = \infty$, and strictly increasing in u. Consequently, (3.14) can be rewritten as follows:

$$\Psi(r, u) = \max_{s} \frac{Y - T(r) - Z(s, u)}{s}. \tag{3.15}$$

It follows from this expression that the equilibrium consumption bundle of a consumer located at r is obtained at the tangency point between the budget line whose slope equals $\Psi(r, u)$ and the indifference curve of level u in the positive orthant of the (z, s)-plane, as illustrated in Figure 3.4. For each r at which the net income $Y - T(r)$ is positive, the unique solution to (3.15) is denoted by $S(r, u)$.

The price of the composite good being 1, the indirect utility when the land rent is R and the net income I is denoted $V(R, I)$. By definition of the bid rent, we have the identity

$$u \equiv V[\Psi(r, u), Y - T(r)]. \tag{3.16}$$

Denoting by $\hat{s}(R, I)$ the Marshallian demand for land and by $\tilde{s}(R, u)$ the Hicksian demand for land, we readily obtain the following well-known identities:

$$S(r, u) \equiv \hat{s}[\Psi(r, u), Y - T(r)] \equiv \tilde{s}[\Psi(r, u), u]. \tag{3.17}$$

We are now prepared to characterize the bid rent and the lot size functions. Applying the envelope theorem to (3.15), we obtain

$$\begin{aligned} \frac{\partial \Psi(r, u)}{\partial r} &= -\frac{T'(r)}{S(r, u)} < 0 \\ \frac{\partial \Psi(r, u)}{\partial u} &= -\frac{1}{S(r, u)} \frac{\partial Z(s, u)}{\partial u} < 0 \end{aligned} \tag{3.18}$$

because T is strictly increasing in r and $Z(s, u)$ strictly increasing in u, whereas using (3.17) yields

$$\frac{\partial S(r,u)}{\partial r} = \frac{\partial \tilde{s}}{\partial R}\frac{\partial \Psi(r,u)}{\partial r} = -\frac{\partial \tilde{s}}{\partial R}\frac{T'(r)}{S(r,u)} > 0 \tag{3.19}$$

because the Hicksian demand for land is always strictly decreasing in rent, and

$$\frac{\partial S(r,u)}{\partial u} = \frac{\partial \tilde{s}}{\partial R}\frac{\partial \Psi(r,u)}{\partial u} > 0$$

because the Marshallian demand for land is strictly decreasing in rent owing to the normality of land. Thus, we have shown:

Proposition 3.4 *The bid rent function is continuously decreasing in both r and u (until it becomes zero). Furthermore, the lot size function is continuously increasing in both r and u.*

When $T(r)$ is linear or concave in distance, $\Psi(r, u)$ is strictly convex in r, as shown by differentiating (3.18) with respect to r and using (3.19).

We now turn to the description of the equilibrium conditions for the monocentric city with N homogeneous consumers, each having a given income Y; landowners are assumed to be absentee. The equilibrium utility u^* is the maximum utility attainable in the city under the market land rent $R^*(r)$. Using (3.16), we obtain

$$u^* = \max_r V[R^*(r), Y - T(r)], \tag{3.20}$$

which is the common utility level in equilibrium.

If one differentiates (3.16) with respect to r and uses Roy's identity, the utility-maximizing choice of a location by a consumer at the residential equilibrium implies

$$S(r, u^*)\frac{dR^*(r)}{dr} + \frac{dT(r)}{dr} = 0. \tag{3.21}$$

That is, at the residential equilibrium, changes in land costs evaluated at the utility-maximizing land consumption are balanced by the corresponding changes in commuting costs. In particular, when the lot size is fixed ($s = 1$), (3.21) becomes $dR^*(r)/dr + dT(r)/dr = 0$, and thus

$$R^*(r) + T(r) = \text{constant}.$$

In this case, the shape of the land rent is the opposite of the shape of the commuting cost function, whereas the consumption of the composite good is the same across consumers. If commuting costs are linear in distance, then the aggregate differential land rent is just equal to total commuting costs when the city is linear, whereas it equals half the total commuting costs when the city is circular. Unexpectedly, these results still hold when the lot size is variable and the utility function well-behaved (Arnott 1979; 1981).

Let $n(r)$ be the consumer density at distance r in equilibrium. Then, we have

$$R^*(r) = \Psi(r, u^*) \quad \text{if} \quad n(r) > 0.$$

If it is assumed that land not occupied by consumers is used for agriculture yielding a constant rent $R_A \geq 0$, the city fringe arises at distance r^* such that

$$\Psi(r^*, u^*) = R_A. \tag{3.22}$$

The bid rent being decreasing in r by Proposition 3.4, the residential area is given by a disk centered at the CBD having radius r^*. As a consequence, the market land rent is given by

$$R^*(r) = \begin{cases} \Psi(r, u^*) & \text{for } r \leq r^* \\ R_A & \text{for } r \geq r^*. \end{cases} \tag{3.23}$$

Because no land is vacant within the urban fringe, we must have

$$n(r) = 2\pi r / S(r, u^*) \quad \text{for all } r \leq r^*, \tag{3.24}$$

and thus the total population N within the urban area must satisfy

$$\int_0^{r^*} \frac{2\pi r}{S(r, u^*)} dr = N. \tag{3.25}$$

In summary, the *residential equilibrium* is described by $R^*(r), n^*(r), u^*$, and r^* satisfying conditions (3.22) through (3.25). Under the preceding assumptions about preferences, income, and commuting costs, the existence of a unique residential equilibrium can be shown to hold (Fujita 1989; Proposition 3.1).

Proposition 3.4 and (3.23) imply that, within the urban area, the market land rent is decreasing as one moves away from the CBD, which is a result that also holds in the Thünian model. Denoting the population density at r by $\delta(r) \equiv n^*(r)/2\pi r$, we see from (3.24) that

$$\delta(r) = 1/S(r, u^*).$$

We may then conclude from Proposition 3.4 that *the equilibrium population density is decreasing from the* CBD *to the urban fringe*, whereas the equilibrium land consumption simultaneously rises. In other words, consumers trade more (less) space for housing against a lower (higher) accessibility to the CBD in a way that allows them to reach the same highest utility level across locations. In monetary terms, a consumer paying a high (low) price for land bears low (high) commuting costs, but the compensation is not necessarily exact because the consumption of the composite good also changes with r. Indeed, each consumer residing further away from the city center has a larger consumption of land and a smaller consumption of the composite good.[13]

Therefore, the equilibrium city accommodating a population of N consumers is described by a circular area centered at the CBD. The consumer density as well as the land rent fall as the distance to the city center rises. This provides an explanation for the fairly general empirical fact that the population density is higher near the city center (where housing costs are high) than at the city outskirts (where housing costs are low). In addition, the size of the residential area depends on the opportunity cost of land but also on the number of consumers, their income, and the value of their commuting costs to the CBD. These relations will be used in Section 3.3.2 to explain another major fact about urban areas, namely suburbanization.

3.3.2 Comparative Statics of the Residential Equilibrium

We can now perform some comparative statics that will shed additional light on real world issues (see Fujita 1989, chap. 3 for more details). First, an increase in the population size has fairly straightforward effects. Indeed, a rising population makes competition for land fiercer, which in turn leads to an increase in land rent everywhere and pushes the urban fringe outward. This corresponds to a well-documented fact stressed by economic historians. Examples include the growth of cities in Europe in the twelfth and nineteenth centuries as well as in North America and Japan in the twentieth century or since the 1960s in Third World countries. All were caused by demographic expansion and rural–urban migrations resulting from technological progress in agriculture, which freed some population from agricultural activity (Bairoch, 1985, chaps. 10 and 14).

We now investigate the impact of a rise in consumers' income Y. Using (3.22) and (3.25), we can readily verify that the residential area expands because the urban fringe moves outward. Although all consumers are clearly strictly better off, the impact on the land rent and the population density is less obvious. An increase in consumers' income raises demand for land everywhere. However, it also leads to a decrease in the relative value of commuting costs, thus making locations in the suburbs more desirable than before the income rise. Consequently, because enough land is available in the suburbs (recall that the additional land available between r and $r + dr$ is $2\pi r dr$), a substantial segment of the population will move from the center to the suburbs. This will in turn decrease the land rent and the population density near the CBD but increase them in the suburbs. In other words, both the land rent and the population density become flatter. Because the locational decision of a consumer is governed by his net income, decreasing the commuting costs has exactly the same impact as increasing Y. We may then conclude that, since the development of modern transportation means (mass transportation and cars) that have followed the Industrial Revolution, income has increased and commuting costs have decreased, generating both suburbanization and a flattening of the urban population densities in many American and European cities (Bairoch 1985, chap. 19).

Finally, consider an increase in the opportunity cost of land as measured by the agricultural land rent R_A. Using (3.22) and (3.25), one can show that the urban fringe shrinks, whereas the equilibrium utility level falls as R_A rises. Then, Proposition 3.4 implies that both the market land rent and consumer density are higher at any distance within the new urban fringe. Hence a higher opportunity cost of land leads to a more compact city with more consumers at each location paying a higher land rent. Increasing the opportunity cost of land therefore leads to more concentrated populations and less well-being for consumers, as suggested by the current situation in many cities in Japan or other countries in East Asia. A high opportunity cost for land may be due to the relative scarcity of land, but it may also find its origin in public policies that maintain the prices of agricultural products far above the international level.[14] This also explains why, for centuries, the spatial extension of towns was limited by returns in agricultural activities as well as by the transport means available to ship produce (Bairoch 1985, chap. 1).

3.3.3 Efficiency of the Residential Equilibrium

It remains to discuss the efficiency of the residential equilibrium. Because this equilibrium is competitive (consumers are price takers) and no externalities are involved, the first welfare theorem proven in Chapter 2 suggests that the equilibrium is efficient. However, we have here a continuum of commodities (land), and thus we need a more specific argument.

It is well known in urban economics that using a utilitarian welfare function leads to the unequal treatment of equals (Mirrlees 1972), whereas equals are equally treated in equilibrium. Such a difference is unexpected, and one might think that competition for space leads to strong social inefficiencies even though our economy is competitive. However, Wildasin (1986a) has shown that this pseudo-paradox arises because the marginal utility of income is different across consumers at different locations. Using a utilitarian approach is therefore unjustified. This fact invites us to consider an alternative approach in which the utility level is fixed across identical consumers.

Assume, then, that all consumers achieve the equilibrium utility level u^* and check whether another feasible allocation $(n(r), z(r), s(r);\ 0 \leq r \leq \hat{r})$ exists that sustains u^* and reduces the social cost C. Note that such an allocation maximizes a *Rawlsian welfare function* (maximizing the minimum utility level in the economy) when the social planner cannot use lump-sum transfers. This has major implications that will be discussed in subsequent chapters.

In our model, the social cost for N consumers to enjoy the utility level u^* is given by the residential cost obtained by summing the commuting costs, the composite good cost, and the opportunity land cost borne by society for this to be possible. Let $Z(s(r), u^*)$ be the quantity of the composite good for which $U[Z(s(r), u^*), s(r)] = u^*$. In consequence, we want to minimize the function

$$C = \int_0^{\hat{r}} [T(r) + Z(s(r), u^*) + R_A s(r)] n(r) dr \tag{3.26}$$

subject to the land constraint

$$s(r)n(r) = 2\pi r \quad \text{for all } r \leq \hat{r} \tag{3.27}$$

and the population constraint

$$\int_0^{\hat{r}} n(r)dr = N. \tag{3.28}$$

Using (3.27) and (3.28), we can readily verify that minimizing (3.26) amounts to solving the following maximization problem:

$$\max_{r,s(r)} S = 2\pi \int_0^{\hat{r}} \left[\frac{Y - T(r) - Z(s(r), u^*)}{s(r)} - R_{\mathrm{A}} \right] r dr$$

subject to (3.28) in which $n(r) = 2\pi r/s(r)$ and Y the fixed income assumed in the equilibrium model in Section 3.3.1.

Neglecting for the moment the population constraint, we may solve this problem by maximizing $[Y - T(r) - Z(s(r), u^*)]/s(r)$ with respect to $s(r)$ at each $r \leq \hat{r}$. By definition of $S(r, u^*)$, it follows that the efficient land consumption $s(r)$ is identical to the equilibrium land consumption of land for each $r \leq \hat{r}$, a condition that holds if and only if

$$\frac{Y - T(r) - Z(s(r), u^*)}{s(r)} = \Psi(r, u^*) \quad \text{for all } r \leq \hat{r},$$

where $\Psi(r, u^*)$ is the bid rent given by (3.15). Therefore, in order to maximize S, $\hat{r}$ must satisfy

$$\Psi(\hat{r}, u^*) = R_A$$

because $\Psi(r, u^*)$ is decreasing in r. Because this equation has a unique solution, it follows that $\hat{r} = r^*$. Given (3.25), it is easily seen that $(s(r), \hat{r})$ satisfies the population constraint (3.28) because $s(r) = S(r, u^*)$ and $\hat{r} = r^*$. Consequently, we may conclude as follows:

Proposition 3.5 *The residential equilibrium is efficient.*

3.3.4 The Case of Multiple Income Classes

In the Thünian model, we have seen that the presence of intermediate goods gives rise to two types of configurations, segregated or integrated (Section 3.2.2). A related concern in understanding the working of a city is to determine how consumers with different incomes organize themselves within the city. Clearly, we will not observe an integrated configuration because consumers endowed with different incomes have different bid rent functions and there is no direct interaction among consumers (e.g., home services from the poor to the rich). Hence, the residential equilibrium involves segregation. What remains to be determined, however, is the shape of the corresponding social stratification at the residential equilibrium.

Consider the case of a finite number m of income classes with N_i consumers in class i; without loss of generality, $Y_1 < Y_2 < \cdots < Y_m$. All consumers have the same utility U and face the same commuting costs $T(r)$. When the utility level of a consumer of class i is equal to u_i, his bid rent function is defined as follows (see (3.15)):

$$\Psi_i(r, u_i) = \max_{s_i} \frac{Y_i - T(r) - Z(s_i, u_i)}{s_i},$$

where $Z(s, u_i)$ is the quantity of the composite good that solves $U(z, s) = u_i$. A residential equilibrium with m income classes is then defined as in the preceding section where the conditions similar to (3.22) through (3.25) are stated for each class i.

As in the Thünian model, *each location is occupied by the consumers with the highest bid rent*. Consequently, the social stratification results from the ranking of the bid rent functions in terms of their slope in a sense that will now be defined. It is apparent that the equilibrium utility levels are such that $u_1^* < u_2^* < \cdots < u_m^*$ because the utility function does not exhibit satiation. Wherever two bid rent curves $\Psi_i(r, u_i^*)$ and $\Psi_j(r, u_j^*)$ with $i < j$ intersect at $\bar{r} \geq 0$, (3.17) and the normality of land imply that the corresponding lot sizes are such that

$$\begin{aligned} S_i(\bar{r}, u_i^*) &\equiv \hat{s}[\Psi_i(\bar{r}, u_i^*), Y_i - T(\bar{r})] \\ &< \hat{s}[\Psi_j(\bar{r}, u_j^*), Y_j - T(\bar{r})] \equiv S_j(\bar{r}, u_j^*). \end{aligned}$$

Hence, by (3.18), $\Psi_i(r, u_i^*)$ turns out to be steeper than $\Psi_j(r, u_j^*)$ at $\bar{r}$. This means that consumers of class i (j) will outbid those of class j (i) on the left (right) side of $\bar{r}$.[15]

Repeating the same argument for each pair (i, j) of income classes, we find that the N_1 consumers of the lowest income class occupy a disk of land centered at the CBD, the N_2 consumers with the second lowest income occupy a ring surrounding this disk, . . . , and the N_m consumers belonging to the richest class are situated in the outermost ring. We thus have the following:

Proposition 3.6 *Assume that consumers have the same preferences and commuting cost function. Then, the social stratification of consumers within the city obeys the rule of concentric rings such that the consumer classes are ranked by increasing income as the distance from the CBD rises.*

Despite strong simplifying assumptions, this result sheds light on the stylized fact, in many U.S. cities, that the poor live near the city center and the wealthy in the suburbs (Wheaton 1977).

Proposition 3.6 also offers a new perspective into the political economy of the city. An increase in the income of the rich consumers relaxes competition for land because these consumers move farther away from the center, making all income groups better off. On the other hand, raising the income of the poor consumers intensifies competition for land and pushes the rich farther away in the suburbs; eventually the poor people are better off but the rich ones are worse off. This suggests a potential conflict between the two classes: the poor have no objection to the rich class becoming richer, but the latter may find it better to keep the poor class poor. This seems to agree with the fact that shocks in the income distribution induce the development of particular urban sections at the expense of others, whereas the rich class members often try to lobby urban governments to implement restrictive zoning policies.

Having said this, we must acknowledge that this proposition is far from providing a complete answer to the stratification problem. Neglected factors governing the distribution of consumers over the urban space include the size of the family, the value of commuting time, the existence of historical and natural amenities, and the financial support of the school system. Although we will not study these factors exhaustively, their impact can be summarized as follows:

1. A larger family has a stronger preference for space, which makes it live farther away from the CBD to benefit from the lower land rent prevailing there (Beckmann 1973).
2. If higher income workers place a higher value on their commuting time, they face a trade-off between a higher land demand (due to normality of land) and the extra value of commuting time. As a result, the low-income consumers reside near the center and the middle class consumers in the suburbs; however, now the high-salary professionals and working couples choose to reside close to the CBD, because of their high value of time, in an urban section different from that of the poor consumers (Fujita 1989, chap. 2).
3. The existence of a well-preserved historical center may lead the rich households to cluster nearby to enjoy the benefit of a rich cultural life (as in Kyoto or Paris). Likewise, natural amenities available near the city limits may induce a similar clustering at the city fringe to permit the rich consumers to benefit from a better natural environment. To the extent that rich households value being together (a club effect), historical or natural amenities may act as a focal point (Bruekner, Thisse and Zenou 1999).
4. When the financing of education is decentralized, families valuing more education (who are often those with higher incomes) similarly cluster in order to supply a better education to their offsprings. This results in higher human capital in the corresponding neighborhoods, thus perpetuating social and spatial segregation (Bénabou 1994; 1996).

3.3.5 Discrete Foundations of Continuous Land Use Models

The neoclassical urban model used so far differs from standard microeconomics in that all the unknowns are described by *density functions*. Instead, one might want to develop a discrete model with a finite number of households each consuming a positive amount of the composite good as well as a positive amount of land. Though Alonso himself has proposed two alternative formulations of such a discrete model, very little work has been devoted to this issue.[16] In this section, we follow Asami, Fujita, and Smith (1990) as well as Berliant and Fujita (1992) and study a simple one-dimensional model in which consumers are homogeneous.

Space is described by the interval $X = [0, \infty)$ with a unit density of land everywhere and the CBD located at the origin; the opportunity cost of land R_A is positive. A finite number n of consumers may be accommodated in this area. The utility of a consumer is $U(z, s)$, where z is the quantity of the composite good, but now $s > 0$ is the size of a lot defined by an interval $[r, r + s) \subset X$. If a consumer occupies the lot $[r, r + s)$, r describes his location, and the commuting cost is defined by tr, where t is a positive constant. All consumers have the same income Y and the same utility function U, which satisfies all the properties stated in Section 3.3.1. As in Section 3.3.1, let $Z(s, u)$ be the positive quantity of the composite good that yields utility level u when a consumer occupies a lot of size $s > 0$. Recall that $Z(s, u)$ is strictly decreasing, strictly convex in s, and such that $\lim_{s\to 0} Z(s, u) = \infty$.

An allocation $(z_i, s_i, r_i;\ i = 1, \dots, n)$ is defined by a consumption bundle and a location for each consumer. It is feasible if and only if no pair of lots overlap. Without loss of generality, we may rank consumers such that $r_1 < r_2 < \cdots < r_n$.

Let $R(r)$ be the land price function defined on X such that a consumer choosing a lot $[r, r + s)$ pays $R(r)s$ for the lot. Then the consumer problem is given by

$$\max_{r,z,s} U(z, s) \quad \text{s.t.} \quad z + R(r)s = Y - tr.$$

This is formally identical to (3.13), where $T(r) = tr$. Therefore, if this consumer chooses location r and achieves the utility level u, then he must choose a lot size $S(r, u)$ that maximizes the bid rent function (3.15).

A residential equilibrium with n consumers is given by a utility level u^* and a land price function $R^*(r)$ together with a feasible allocation $(z_i^*, s_i^*, r_i^*;\ i = 1, \dots, n)$ such that the following conditions hold:

$$R^*(r) \geq \max\{\Psi(r, u^*), R_A\} \tag{3.29}$$

$$R^*(r_i^*) = \Psi(r_i^*, u^*) \quad i = 1, \dots, n \tag{3.30}$$

$$R^*(r_n^*) = R_A \tag{3.31}$$

$$s_i^* = S(r_i^*, u^*) \quad i = 1, \dots, n \tag{3.32}$$

$$r_1^* = 0 \quad \text{and} \quad r_{i+1}^* = r_i^* + s_i^* \quad i = 1, \dots, n-1, \tag{3.33}$$

where the bid rent function $\Psi(r, u^*)$ is defined by (3.15) for all $r \geq 0$. The condition (3.31) on the land rent for the last consumer allows us to avoid unnecessary technical difficulties,[17] whereas the last condition states that there is no vacant land within the city.

Clearly, because the bid rent function decreases with distance, the equilibrium rents R_i^* satisfy $R_1^* > R_2^* > \cdots > R_n^* = R_A$. Furthermore, since $S(r, u^*)$ is strictly decreasing in r (Proposition 3.4 (ii)) $i < j$ implies that $s_i^* < s_j^*$. So, we have shown

Proposition 3.7 *Consider any finite number of consumers and assume that a residential equilibrium exists.*[18] *Then, this residential equilibrium is such that the land rent decreases as one moves away from the CBD whereas consumers with larger lots locate farther from the CBD than consumers with smaller lots.*

This means that the residential equilibrium with a finite number of consumers displays the same basic features as the continuous standard model of urban economics. However, the preceding discrete model suffers from a serious defect, namely, each consumer pays the same price for each unit of his lot, and thus the landowner may want to extract more from the consumer or a consumer may buy more land for resale to the next one. To avoid this difficulty, one may either assume that arbitrage is prohibitively costly or that a consumer located at r pays a price given by

$$\int_r^{r+s} R(y)dy$$

for the lot $[r, r + s)$, which is also suggested by Alonso. Proposition 3.7 remains essentially the same in this alternative model, but the analysis is more complex (Berliant and Fujita 1992).

Note, finally, that Asami, Fujita, and Smith (1990) have shown that the standard continuous model provides a good approximation of the discrete model considered in this section when n is large enough. In particular, a finite economy with n consumers and a continuous economy with a mass N of consumers each have a cumulative population distribution; these authors show (Theorem 5) that the two sequences of normalized (by the population size) distributions have the same limit as $n = N \to \infty$.

3.3.6 Notes on the Literature

The urban land use model presented in Section 3.3.1 is essentially a simple version of neoclassical urban models, which were developed by Beckmann (1957; 1969), Alonso (1960; 1964), Muth (1961; 1969), Mills (1967; 1972b), Casetti (1971), and Solow (1973). However, it seems fair to say that Beckmann's short article published in the first issue of the *Journal of Economic Theory* and based on his 1957 discussion paper is really pathbreaking in that it is not only a concise statement of the standard monocentric model but also a precursor to several later contributions.

Definitions of closed and open cities were introduced by Wheaton (1974), whereas the public ownership model is credited to Solow (1973). The existence of a residential equilibrium with a continuum of locations and a single marketplace for heterogeneous consumers has been established by Fujita and Smith (1987). The study of the monocentric model when consumers have heterogeneous tastes as described by the logit can be found in Anas (1990), whereas the extension of the standard model to several prespecified centers was considered by Papageorgiou and Casetti (1971).

The comparative statics of the residential equilibrium was first studied by Wheaton (1974) in the case of homogeneous consumers. The optimal city was studied by Mirrlees (1972), who used a utilitarian welfare function. The approach taken in Section 3.3.3 is based on Herbert and Stevens (1970), who retained a discrete space and used the duality theorem of linear programming. The general analysis of the residential equilibrium with several income classes has been presented by Hartwick, Schweizer, and Varaiya (1976).

Finally, the discrete foundations of the continuous urban model have been criticized by Berliant (1985). Possible solutions have been investigated by Asami, et al. (1990), Papageorgiou and Pines (1990), and Berliant and Fujita (1992).

The state of the art in urban economics is summarized in the two complementary books by Fujita (1989) and Papageorgiou and Pines (1999), whereas a historical and methodological outlook of this field is provided by Baumont and Huriot (2000).

3.4 CONCLUDING REMARKS

We have seen how land use patterns and land rent profiles can be determined in competitive land markets once it is assumed that a center exists where (some or all) tradable goods have to be shipped. As in Chapter 2, we have assumed that there are no physical differences in land at different locations. The differences in land rents can therefore be attributed to the relative advantage of each location compared with the extensive margin of land use. In this sense, the land rent corresponds to a *locational rent*. This concept of rent is to be contrasted to the more standard concept of scarcity rent, which could be integrated into the Thünian model by assuming that the “isolated state” is replaced by a “small circular island.” At the border of the island, the land rent would be positive, and this value would express the global scarcity of land, whereas the difference in the value of the land rent inside the island would still have the nature of a locational rent. In the Thünian model, the land rent is equal to the excess of revenues obtained from the sale of goods produced by using land over payments to nonland factors used in production and transportation. This is why the bid rent function is obtained from a condition of zero profits, which can be interpreted as a free-entry condition of producers in each activity considered.

When consumers (instead of producers) use land, the mechanism leading to the formation of the bid rents is similar to the one uncovered by Thünen provided that the utility level is given by the reservation utility and the population size is variable (this amounts to a condition of free entry). This is called the *open city* model, an example of which will be discussed in the next chapter. By contrast, in the *closed city* model in which the population size is fixed, the utility level is endogenous. This requires a more general approach to the formation of bid rents such as that studied in Section 3.3.1. Both types of models (closed city and open city) are useful because they correspond to different situations and lead to similar results.

In addition, we have implicitly assumed absentee landlords. That is, the land rent earned within the city goes to landlords who do not reside within the city; hence, the rent does not feed back into consumers' incomes. Both the closed city and open city models can be extended to cope with public ownership of land in which the aggregate land rent is first collected by a public agency and then equally shared among consumers. The analysis remains essentially the same. The choice of a particular specification (open versus closed, absentee landlords versus public property) is dictated by the main features of the problem under consideration. In this chapter, we have chosen to present the most popular model, and we refer to Fujita (1989) for more details regarding the other approaches.

NOTES

1. Either approach will be used in this book. Roughly speaking, we can say that the former is followed in models with a finite number of agents to locate, whereas the latter will be encountered in models with a continuum of agents.
2. All these assumptions are to be contrasted to those made in the quadratic assignment problem discussed in Section 2.2.
3. In this respect, the books by Bairoch (1985) and by Hohenberg and Lees (1985) offer both a great deal of relevant information.
4. In general, a point is described by its radius and its angle, but we may omit the angle because space is featureless around the city.
5. As in Lucas (2001), we treat here a unit of land as a given combination of land and labor. Alternatively, we may consider that p_i introduced below represents the crop i's price net of all input-costs other than land rent. The cost of labor is explicitly accounted for in Section 3.2.3.
6. It is worth noting that Thünen used a more general specification of the transport cost involving two components. The first component corresponds to a monetary cost proportional to the quantity shipped and the distance covered (like ours), whereas the second is given by a fraction of the initial shipment's melting during the transport. For example, Thünen supposed that the cost of shipping grain consists partially of the grain consumed on the way by the horses pulling the load. This anticipates the iceberg cost used by Samuelson (1954a, 1983), Nerlove and Sadka (1991), and Krugman (1991a,b).

7. If the number of locations were finite, the land rent would be given by the outcome of an English auction in which the commodity is sold at the second highest reservation price. When the distance between adjacent locations along any ray goes to zero, the second highest reservation price tends to the highest reservation price at each location as given by Proposition 3.1 (Asami 1990). However, we must stress that, in more general settings in which the land use pattern is determined together with prices – wages or utility levels, say – this is no longer true. In such contexts, it is not clear how the bid rent function may emerge from a standard auctioning process.
8. Given (3.3), it is readily verified that the total surplus is identical to the *aggregate land rent* in this model.
9. See Chapter 6 for a similar result in a different context.
10. The same crop, however, may be raised within two different zones.
11. To focus on lot size and population density changes within the city, we use a simple utility with two arguments, z and s. However, the model can readily be extended to the case of several consumption goods as well as to nonland input for housing (see Fujita 1989, 44).
12. If the net income $Y - T(r)$ is positive, the utility-maximizing bundle (z, s) exists and is unique because U is strictly quasi-concave.
13. As seen in the foregoing, there is exact compensation when the lot size is fixed.
14. According to Ohmae (1995, 48),

    within a 50-kilometer radius of Tokyo, 65 percent of land – nearly 330,000 hectares of some of the most expensive property in the world – is devoted to wildy inefficient agriculture. If only one quarter of this land were sold for private housing, Tokyo-area families would be able to afford 120 to 150 square meters of living space, instead of today's average of 88 square meters. Moreover, cheaper – and more available – land would cut the cost of essential public work like providing better sewage, removing traffic bottlenecks, and double-tracking commuter trains.

15. Note that the lot size is discontinuous at the border between two adjacent social areas. This corresponds to the discontinuity observed in the employment level at the border between two adjacent zones of production in the neoclassical model of land use.
16. By contrast, the connection between the Arrow–Debreu model and the continuous approach developed by Aumann has attracted much attention (Hildenbrand 1974).
17. When this condition is replaced by the inequalities $\Psi(r_n, u^*) \geq R_A$ and $\Psi(r_n + s_n, u^*) \leq R_A$, there is a continuum of equilibria (see Asami et al. 1990, Theorem 2).
18. Standard tools of general equilibrium analysis are not applicable here because of major nonconvexities. However, existence, uniqueness, and optimality have been shown by Asami et al. (1990).

4

Increasing Returns and Transport Costs:

The Fundamental Trade-Off of a Spatial Economy

4.1 INTRODUCTION

As seen in Chapter 2, a homogeneous space together with the competitive mechanism is not compatible with the existence of economic agglomerations such as cities. This is why we broached the assumption of homogeneity in Chapter 3 by assuming the existence of a prespecified center to which workers must commute. We did not, however, provide any justification for the existence of such a center.

One of the main reasons for a city center is the presence of scale economies in the aggregate, but such social returns remain to be explained. The standard approach is to appeal to Marshallian externalities. As argued in Chapter 1, this often amounts to using a black box. More interesting is a recent and growing literature that explores the specific factors that can generate such social scale economies. In this chapter, it is not our intention to provide a full survey of this new literature. Instead, we want to stress the importance of two forces, namely, the *diversity of intermediate goods* and the effect of the *matching process on the labor market*. Each factor lies at the origin of the division of labor and provides a foundation for increasing returns in the aggregate. Specifically, we will illustrate the advantages of specialization by showing how increasing returns may arise in the final goods sector when the intermediate goods sector is described by a monopolistic competition model. In addition, we will show how imperfect competition in thick labor markets, such as those encountered in big cities, allows for a reduction in average matching costs. This in turn results in higher wages. Both models illustrate how a growing urban population has permitted gains to be generated from both specialization and matching. In this sense, they are very much in the spirit of Adam Smith ([1776] 1965, 17) when he wrote the following:

> There are some sorts of industry, even of the lowest kind, which can be carried on no where but in a great town. A porter, for example, can find employment and subsistence in no other place. A village is by much too narrow a sphere for him; even an ordinary market town is scarce large enough to afford him constant occupation.

In fact, the degree of increasing returns may be so high that a single firm operates at the CBD. Such a situation is reminiscent of the "factory town" in which workers are hired by a single firm that also holds the land leased to its workers. Besides production activity, the firm therefore assumes the tasks of a *developer* (Mirrlees 1972; Henderson 1974). In such a context, land and the composite good may be viewed as "intermediate inputs" used by the firm to attract workers who move in to produce the final goods in the firm and to live in the city. In this sense, it is fair to say that the factory town resembles a vertically integrated structure. Furthermore, this approach allows one to regard a city as a firm maximizing some objective. Specifically, the land-development company, owned by Arrow–Debreu shareholders, seeks to maximize land rents net of any cost.

What makes this setting interesting for our purpose is that the firm must take actions to attract workers from the rest of the economy, which has powerful implications. Indeed, the firm must meet two requirements. First, it must pay the migrants a wage high enough to compensate them for the extra costs they incur while residing in the city (land rent plus commuting costs). Second, the firm must give the workers a utility level at least as high as the best alternative they can secure in the rest of the economy, which means that firms behave as *utility-takers*. In addition, it is possible to describe an entry–exit process similar to that encountered in Marshallian theory of competition in which the free-entry, zero-profit equilibrium involves a system of factory towns, each accommodating the optimum population of workers. At this equilibrium, there are positive transport costs within cities. Note that this result does not contradict the spatial impossibility theorem because firms behave as utility-takers and not as wage-takers.

There are two main results. The first one states that, at the zero-profit equilibrium, each firm must select an output, or equivalently an employment level, arising in the domain where it faces increasing returns. Roughly speaking, the essence of the argument is as follows. The average total (production plus transport) cost of each firm must be minimized at the long-run output and is, therefore, equal to the corresponding marginal total cost. Furthermore, because increasing the number of workers causes transportation costs to rise more than proportionally, the marginal transport cost exceeds the average transport cost. Hence, the marginal production cost of each firm must be lower than its average production cost. This means that the optimal firm size must be in the phase of increasing returns.[1]

The second result has been described by Serk-Hanssen (1969) and Starrett (1974). Although workers are paid at their marginal productivity, increasing returns do not prevent the emergence of a first best optimum. This is so because the losses incurred by each firm are exactly compensated by the aggregated differential rent within each city at the zero-profit equilibrium. Indeed, when firms (or, more generally, any group of agents forming a production coalition) capitalize the land rent they create by their activity, everything works as if firms were able to capture the whole consumer surplus, whence firms do the socially optimal things when maximizing profit, as they do under perfect price discrimination (Spence 1976). This should not come as a surprise, for we have seen that a competitive land market leads residents to pay the highest surplus they can afford in order to occupy a particular location (see Section 3.3.1); thus, land capitalization can be viewed as way of extracting the whole consumer surplus. Or, as Vickrey (1977, 343) put it in a very neat and transparent way:

> Urban land rents are, fundamentally, a reflection of the economies of scale of the activities that are carried on within the city, and that efficient organization of a city, or even of the urban life of a nation as a whole, requires that these land rents, or their equivalent, be devoted primarily to the financing of the intramarginal residues that represent the difference between revenues derived from prices set at marginal costs and the total cost of the activities characterized by increasing returns.

This is important because the presence of increasing returns is known to make it very unlikely for producers to respond optimally to a central message (e.g., a price system in a competitive setting). Here, optimality is obtained by introducing three main changes in the competitive model: (1) firms are wage-makers (instead of wage-takers); (2) they are utility-takers, that is, firms do not assume they can manipulate the utility level of consumers; and (3) they are able to capitalize the differential land rent they create into their payoffs. Though the first two assumptions are not related to space per se, the third one does require an explicit accounting for land and transportation.

These results are probably the most distinctive contribution of location theory to economics because they show how spatial friction costs go hand in hand with the presence of increasing returns in production as well as how site rents play an essential role in the emergence of the optimum in an economy involving increasing returns in production.

From the spatial point of view, these results confirm Mills' (1967) view that *cities form in the economy because there are scale economies in production.* Nevertheless, despite the presence of increasing returns in production, the profit-maximizing population of workers is finite, and the corresponding urban area is bounded. It turns out, therefore, that scale economies in production are damped by scale diseconomies arising in transportation (even when the individual commuting cost displays long-haul economies). As also acknowledged by Mills (1967), the size of a city is determined by a trade-off between increasing returns and transportation costs. Stated differently, in the absence of scale economies in production, there would be no city (backyard capitalism), whereas, with no transportation costs, there would be a single city in the economy (the world megalopolis). Thus, it is not too much to say that this trade-off is central to the understanding of a city.

An alternative institutional system is obtained by assuming that a group of workers decides to form a cooperative producing the final good. They do so because they understand that, by joining efforts, they may enjoy the benefits of increasing returns (at least up to some employment level). To this end, they cluster within a small area and form a *community*, which then establishes a local government. This government is entitled with the objective of maximizing the well-being of the community members by choosing the optimum community size. This problem is related to the previous one in that each community's behavior encapsulates a developer's problem, and competition among communities yields the same long-run equilibrium. In fact, both approaches amount to assuming a "market for cities" (Henderson) in which cities are created until no opportunity exists for a developer or a community to build a new one.[2] This leads to a simple theory of the urban system, which can in turn be extended to explain why different types of cities emerge (Henderson 1974; Mirrlees 1995).

It remains to consider the classical trade-off between increasing returns at the firm's level and transportation costs, which goes back at least to Kaldor (1935) and Lösch ([1940] 1954). Despite the difference in settings, this trade-off is similar to those mentioned in the foregoing, although its description involves a particular form of imperfect competition known as *spatial competition*.

When production entails increasing returns at the firm's level and demand is spatially dispersed, the economy accommodates only a finite number of firms, which are imperfect competitors because they can derive monopoly power from their geographic isolation. Treading in Hotelling's footsteps, Kaldor argued that space gives this competition a particular form. Because consumers buy from the firm with the lowest "full" price, that is, including the transport cost, each firm competes directly with only a few neighboring firms regardless of the total number of firms in the economy. Or, as Kaldor (1935, 391) put it:

> Looked at from the point of view of any seller, a change of price by any other particular seller (the prices of the rest being assumed as given) is less and less important for him, the further away that particular seller is situated.

Therefore, the process of spatial competition takes place among the few and should be studied within a framework of interactive decision making. This is the other main contribution of location theory to economics in that it shows the importance of strategic considerations in the formation of prices in spatial markets. However, it was ignored until economists became fully aware of the power of game theory for studying competition in modern market economies. Following the outburst of industrial organization theory since the late 1970s, it became natural to study the implications of space for competition. New tools and concepts are now available to revisit and formalize the questions raised by early location theorists such as Launhardt ([1885] 1993), Hotelling (1929), and Lösch ([1940] 1954).

In that context, scale economies in production have another far-reaching implication that is of direct relevance for us: the number of marketplaces open in the space-economy is likely to be suboptimal. Or, to put it differently, spatial markets being incomplete because of nonconvexities in technologies, an equilibrium allocation is generally not efficient. This sheds additional light on the trade-off between increasing returns and transportation costs in a spatial economy in which firms compete strategically to attract consumers. As will be seen, large-scale economies yield a sparse production pattern, whereas high transportation costs lead to a dense spatial configuration of firms. And it is only when both become negligible that the market equilibrium approaches the competitive outcome and so gets close to the optimum.

Spatial competition models can be referred to as "location without land" models because no land market exists. Yet, they can be extended to integrate a land market when it is recognized that consumers are mobile. Introducing a land market and consumer mobility into spatial competition models has at least two important consequences. It permits the location decisions of firms and consumers to be jointly endogenous, and it permits allocation mechanisms based on land capitalization, as in the literature on local public goods. Indeed, a differential land rent is generated by the relative proximity to the stores. When each firm is allowed to collect the extra rent that its activity creates, the discrepancy between the market equilibrium and the optimum vanishes, thus extending the results obtained in the foregoing sections inasmuch as firms now behave strategically.

Although they seem to be very different, the two approaches taken in this chapter are complementary. In urban system models, we consider a framework in which production occurs at some given city centers and determine the corresponding spatial distribution of workers and consumers. In spatial competition models, we suppose a given spatial distribution of demand and ask at which places production will occur.

The organization of the chapter reflects what has been said above. The microeconomic underpinnings for increasing returns at the city level are spelled out in Section 4.2. In Section 4.3, we study the trade-off between scale economies and commuting costs in the context of city formation. The factory town model (Section 4.3.1) and the community model (Section 4.3.2) are successively discussed and compared. These models then serve as a basis for the analysis of urban systems (Section 4.3.3). In Section 4.4, we extend the framework in order to figure out how the interplay between increasing returns and transport costs between cities explains why cities are specialized in the production of a very small number of goods or, instead, are diversified because they nest a wide array of activities. After that, we turn in Section 4.5 to the study of the trade-off between increasing returns and transport costs in the context of spatial competition. Specifically, we first consider the market equilibrium (Section 4.5.1) and then the optimum (Section 4.5.2). The two solutions are different, but we will see how they can be reconciled once firms are allowed to capitalize the differential land rent they create into their profits (Section 4.5.3). Our conclusions are presented in Section 4.6.

4.2 MICROFOUNDATIONS OF INCREASING RETURNS AT THE CITY LEVEL

Recently, alternative (but not exclusive) explanations have emerged that all focus on what is usually called "external scale economies." They turn out to be especially relevant for our purpose because they implicitly consider an urban or industrial agglomeration while yielding similar reduced forms linking wage and productivity to the size of the labor force. Three main lines of research can be distinguished (Duranton and Puga 2000)[3] as follows:

1. The most popular one goes back to Alfred Marshall and deals with the many advantages that firms producing similar goods may exploit by co-locating. For example, different agents own different bits of information and their gathering yields a higher level of knowledge ("the secrets of the industry are in the air"); this in turn improves their productivity. Marshallian externalities were formally introduced in urban economics by Mills (1967) and Henderson (1974) as an externality positively affecting the cost of firms located in a CBD; their intensity depends on the number of firms or on the volume of production. More recently, detailed frameworks have been proposed that allow one to show how the need to exchange information may lead heterogeneous firms (agents) to form a CBD (a cluster). They will be further discussed in Chapters 6 and 8.
2. The Chamberlinian approach rests on the following popular idea: A large market (think of a metropolitan area) allows for a large number of intermediate commodities (Ethier 1982) or final goods (Krugman 1980). In particular, intermediate commodities can be used as inputs to enhance the productivity of the final sector, thus resulting in wages that increase with the size of the urban labor force (Abdel-Rahman and Fujita 1990). And, indeed, the importance for local development of diversified and nontradable inputs (such as legal and communication services, nontraded industrial inputs, maintenance and repair services, finance) is a well-documented fact (Hansen 1990; Saxenian 1994).
3. The third approach is both old and new. It is old because the ideas were clearly spelled out by Adam Smith and further elaborated by Alfred Marshall, but it is new because their ideas were formalized only recently. For our purpose, it is convenient to retain two specific models. In the first, developed by Helsley and Strange (1990), it is shown that a large city allows for a better average match between heterogeneous workers and firms' job requirements.[4] Hence, something like an agglomeration economy is at work when new firms, workers, or both enter the city labor market, thus allowing firms located within the CBD to pay higher wages. In the second, proposed by Duranton (1998), it is argued that a large market permits workers to become more specialized and, therefore, to be more efficient when they are gathered within cities.[5]

In what follows, we have chosen to focus on the Chamberlinian idea of diversity of inputs used in the final sector as well as on the Smith–Marshallian idea that a thick market allows for a better matching between workers and jobs. Clearly, for increasing returns to emerge, some types of indivisibilities must exist within the city. What is shown is that such indivisibilities are expressed by wages that increase with the size of the labor force. Note that another reason for our choice of menu is that these two approaches rely on the two canonical models used in geographical economics, namely, the CES (Section 4.2.1) and spatial competition (Section 4.2.2) models.

4.2.1 A Chamberlinian Approach: The Diversity of Intermediate Inputs

In this section, we illustrate the advantages of diversification in intermediate goods. Assume that the final sector operates under constant returns to scale and perfect competition. Without loss of generality, this sector may then be represented by a single firm whose production function is assumed to be as follows:

$$X = \left\{ \int_0^M [q(i)]^\rho di \right\}^{1/\rho}, \tag{4.1}$$

where ρ takes a value strictly between 0 and 1.[6] In this expression, X is the output of the firm whose price is normalized to one, $q(i)$ the quantity of variety i used, and M the number (or mass) of intermediate goods available in the city.

As observed by Ethier (1982), (4.1) can be interpreted as the production function of a competitive firm that has constant returns with respect to a given number M of specialized inputs $q(i)$.[7] However, this function exhibits increasing returns in the number M of intermediate goods. Suppose, indeed, that each intermediate good is sold at the same price $\bar{p}$, and let E denote the expense of the firm on all the intermediate goods. Then, the consumption of each variety by a firm of the final sector is such that $q(i) = E/M\bar{p}$ for all $i \in [0, M]$. Plugging this expression into (4.1), we obtain

$$X = \frac{EM^{(1-\rho)/\rho}}{\bar{p}}.$$

Hence, for any given value of E, production strictly increases with the mass of intermediate goods as long as $\rho < 1$. The more specialized the intermediate goods, the stronger this effect; that is the smaller ρ. Hence, the working of the final sector depends on the way the intermediate sector operates.

Consider a monocentric city with a final goods industry and an intermediate goods industry. The latter supplies specialized services to the former. The production function of a firm belonging to the final sector is given by (4.1), whereas the production function of the service firms is described below. Both types of firms are located in the CBD.

Because of specialization in production, each variety is produced by a single firm according to an identical technology for which the only input is labor. The total amount of labor required to produce the quantity $q(i)$ of the intermediate good i is given by

$$l(i) = f + cq(i), \tag{4.2}$$

where f is the fixed labor requirement and c the marginal labor requirement. Clearly, this technology exhibits increasing returns to scale.

Let w denote the common wage prevailing in the city. If $p(i)$ denotes the price of the intermediate good i, the representative firm selling the final good chooses $q(i)$ so as to maximize its profit given by

$$X - \int_0^M p(i)q(i)di$$

subject to the production function (4.1). The first-order conditions yield the input demands as

$$q^*(i) = Xp(i)^{-\sigma} P^{\sigma} \quad i \in [0, M], \tag{4.3}$$

and the total expenditure as

$$\int_0^M p(i)q^*(i)di = PX$$

where P is the price index for the intermediate sector defined as follows:

$$P \equiv \left[\int_0^M p(j)^{-(\sigma-1)}dj \right]^{-1/(\sigma-1)}, \tag{4.4}$$

whereas

$$\sigma \equiv \frac{1}{1-\rho},$$

is the elasticity of substitution between any two varieties that varies between 1 and ∞. Because the profit function of the representative firm, given by $X - PX = (1 - P)X$, is linear in X, the equilibrium price index P^* must satisfy

$$P^* = 1 \tag{4.5}$$

for the equilibrium output to be positive and finite.

Herein we follow Abdel-Rahman and Fujita (1990) and assume that the intermediate sector is described by a market structure in which (1) each firm produces one intermediate good (monopolistic) and (2) profits are just sufficient to cover average costs (competition). That is, we use Dixit and Stiglitz' (1977) monopolistic competition model in which the representative consumer is replaced by the representative firm of the final sector. Since there is a continuum of firms in the intermediate sector, each one is negligible in the sense that its action has no impact on the market. Hence, when choosing its price, a firm accurately neglects the impact of its decision over the magnitudes X and P. In addition, because firms sell differentiated intermediate goods, each one has some monopoly power in that it faces an isoelastic demand function, the elasticity of which is σ.

The profit of firm i is

$$p(i)q^*(i) - wl(i).$$

Because demands (4.3) are symmetric and isoelastic, the equilibrium price is the same across firms and is equal to the common marginal production cost times a positive relative markup

$$p^* \equiv p^*(i) = cw/\rho \qquad i \in [0, M] \tag{4.6}$$

common to all firms.

Firms will enter the intermediate goods industry until profits are zero, that is, $p^*q^* - wl^* = 0$. Substituting (4.2) and (4.6) into this equality yields the equilibrium output common to all firms of the intermediate sector as follows:

$$q^* \equiv q^*(i) = \frac{f}{c}\frac{\rho}{1-\rho} \tag{4.7}$$

which in turn yields the equilibrium labor consumption:

$$l^* \equiv l^*(i) = \frac{f}{1-\rho}. \tag{4.8}$$

Assuming that full employment prevails in the city when the size of the labor force is N requires

$$N = M^*l^*. \tag{4.9}$$

Using (4.8), (4.9) may be rewritten so as

$$M^* = (1-\rho)N/f, \tag{4.10}$$

which means that the equilibrium number of intermediate goods is increasing in the labor force but decreasing in its own fixed cost. In other words, the process of specialization is limited by the size of the labor market (N) as well as by the presence of fixed costs in the intermediate sector (f). The equilibrium number of intermediate goods also rises with the degree of product differentiation characterizing this sector (ρ is small).

We now determine the unknowns X and w as functions of the mass of intermediate goods M. Using (4.1) and (4.7), we obtain the following relationships:

$$X = K_1\, M^{1/\rho}, \tag{4.11}$$

where $K_1 \equiv f\rho/(1-\rho)c$; Furthermore, using (4.4)–(4.6) leads to

$$w = K_2\, M^{(1-\rho)/\rho}, \tag{4.12}$$

where $K_2 \equiv \rho/c$. This implies that both the equilibrium output of the final sector and the equilibrium city wage increase with the mass of the intermediate goods.

As a consequence, (4.10) and (4.11) allow us to rewrite the city production function (4.1), built here from individual components, as follows:

$$X = AN^{1/\rho}, \tag{4.13}$$

where $A \equiv f^{-(1-\rho)/\rho}\rho(1-\rho)^{(1-\rho)/\rho}/c$ is a positive constant depending on the key parameters of the economy; in particular, A decreases with both the fixed labor requirement f and the marginal labor requirement c.

Thus, in the aggregate, production in the final sector exhibits increasing returns in the labor force (the exponent of N in (4.13) is greater than 1 when $\rho < 1$) even though the production function in the final sector displays a priori constant returns to scale. This is because the number of specialized firms in the intermediate sector rises with the population size, thus permitting a higher degree of specialization.[8] The concept of city production function is not obvious because the city is not an agent per se but a collection of agents, each with her own interest. Hence, aggregate production functions should be built inside the model. This is precisely what we have achieved in the foregoing.

Depending on the value of ρ, the degree of increasing returns might be high enough for a "large" number of workers to reside within the same city. In particular, when $\rho < 1/2$, the marginal product of labor rises at an increasing rate with the population size. This does not strike us as a realistic outcome, and we find it more reasonable to assume that the increase in the marginal productivity falls with N. Thus, provided that $\rho > 1/2$, the process of specialization cannot be pursued indefinitely. Still, even in this case, the economy displays what may be called "returns to scale at the city level."

The equilibrium wage may be obtained from (4.10) and (4.12)

$$w^* = AN^{(1-\rho)/\rho},$$

which also increases with the size of the labor force – a result that has been empirically tested in several American cities (Rauch 1993; Peri 2001). Indeed, when more workers are available in the city, more firms can enter the intermediate sector, generating higher wages. However, when $\rho > 1/2$, the equilibrium wage rises at a decreasing rate because $1/\rho - 2 < 0$.

As discussed in Section 3.3, an increasing population leads to an expansion of the residential area, which yields in turn higher land rents and longer commuting. Consequently, the city size is determined endogenously at the solution to the trade-off between increasing returns in the intermediate sector and workers' commuting costs to the CBD.[9] Of course, cities specialized in different final goods will have different production functions, yielding different sizes.

Loosely speaking, we may thus conclude as follows: As in Adam Smith, the division of labor within the city is limited by the number of workers but, as in Young (1928), the extent of the labor market is itself limited by the specialization of labor through the degree of increasing returns in the intermediate sector.

4.2.2 A Smith–Marshallian Approach: The Matching Process on the Labor Market

We now follow a different path by assuming a population of workers who are heterogeneous in the type of work they are best suited for. For example, think of a big city (such as London or New York) that accommodates a large number of lawyers with specific training. Once it is recognized that the labor force is heterogeneous in the skill space, it should be clear that firms have incentives to differentiate their technologies in much the same way as firms have incentives to locate at different places when consumers are dispersed over space (see Section 4.5). Indeed, firms are then able to obtain market power in the labor market that allows them to set wages below the productivity of workers. The approach taken here combines ideas from Helsley and Strange (1990) as well as from Hamilton, Thisse, and Zenou (2000).[10]

Consider a monocentric city with M firms located in the CBD. As in the foregoing section, we assume that firms sell their output at a given market price (we take this output as the numéraire). For simplicity, a firm is fully described by the type of worker it needs. Firm i's skill requirement is denoted by r_i $(i = 1, \ldots M)$ in some skill space C. Labor is the only input, and production involves constant returns to scale once some entry cost measured in terms of the numéraire has been paid.

There is a continuum of workers of size N with heterogeneous skills. Workers are heterogeneous in the type of work they are best suited for, but there is no ranking in any sense of these types of work. A worker's skill type, denoted by $r \in C$, is distributed in the skill space. The characteristics of a worker relevant to firms are summarized by her skill. Finally, each worker supplies one unit of labor provided that her wage net of training costs paid by the worker (her earnings) is positive (without loss of generality, the reservation wage is normalized to zero).

Each firm has a specific technology such that workers can produce output only when they match the firm's skill needs perfectly. Because workers are heterogeneous, they have different matches with the firm's job offer. Thus, if firm i hires a worker whose skill differs from r_i, the worker must get trained, and her cost of training to meet the firm's skill requirement is a function of the difference between the worker's skill r and the skill requirements r_i.

When workers are heterogeneous and the market is sufficiently thick, it seems natural to assume that firms cannot identify an individual's skill before employment. Instead, firms know the statistical distribution of individual skills. For simplicity, the skill space C is described by the circumference of a circle that has length L. Individuals' skills are continuously and uniformly distributed along this circumference; the density is constant and denoted by Δ. The density Δ expresses the *thickness* of the labor market, and L measures the *degree of diversity* in workers' skills. When the population of workers is heterogeneous, the extent of the labor market is described by these two parameters. We will see that they play different roles in the market equilibrium. Clearly, L and Δ must be such that $N = L\Delta$. We assume that firms' job requirements r_i are equally spaced along the circumference C so that L/M is the distance between two adjacent firms in the skill space.[11]

The training cost function is $s|r - r_i|$, where $s > 0$ is an inverse measure of a worker's ability to learn how to adjust to a technology different from her skill. After training, all workers are identical from the firm's viewpoint because their ex post productivity is observable and equal to g by convention.

Although firms do not observe workers' types, workers know their own type and observe the firms' skill needs.[12] In order to induce the appropriate set of workers to take jobs with the most suitable firm, workers must pay at least some part of the training cost. In addition, since the labor supply of a worker is inelastic, firms cannot offer a wage menu so that the worker must pay for all the costs of training that are not observable to the firm (hence resolving the adverse selection problem). Consequently, each firm i offers the same wage to all workers conditional on the worker having been trained to the skill r_i. Each worker then compares the wage offers of firms and the required training costs; she simply chooses to work for the firm offering the highest wage net of training costs.

The wage-setting game proceeds as follows. First, firms simultaneously choose their gross wage offers. Workers then observe all wage offers and choose to work for the firm that yields the highest net wage. Since each firm anticipates workers' choices, it will hire all workers who prefer to work for it. In equilibrium, there is no quits or layoffs because both firms and workers have no incentive to deviate. The allocation of workers among firms is based entirely on individual competitive advantage.

Consider firm i. If the firms on each side of it offer wages w_{i-1} and w_{i+1}, respectively, then firm i's labor pool consists of two subsegments whose outer boundaries are $\bar{r}_i$ and $\bar{r}_{i+1}$. The worker at $\bar{r}_i$ receives the same net wage from firm i and firm $i - 1$, whereas the worker at $\bar{r}_{i+1}$ receives the same net wage from firm i and firm $i + 1$. Because firm i knows the training cost function and all firms' job requirements, it can determine $\bar{r}_i$ and $\bar{r}_{i+1}$. Specifically, $\bar{r}_i$ is the solution to the equation: $w_i - s(r_i - \bar{r}_i) = w_{i-1} - s(\bar{r}_i - r_{i-1})$, so that

$$\bar{r}_i = \frac{w_{i-1} - w_i + s(r_i + r_{i-1})}{2s}. \tag{4.14}$$

Firm i attracts workers whose skill type lies in the interval $(\bar{r}_i, r_i]$ because they obtain a higher net wage from firm i than from firm $i - 1$. Workers with skill types in $[r_{i-1}, \bar{r}_i)$ prefer to work for firm $i - 1$. Similarly, we can show that

$$\bar{r}_{i+1} = \frac{w_i - w_{i+1} + s(r_i + r_{i+1})}{2s}. \tag{4.15}$$

Firm i's labor pool thus consists of all workers with skill types in the interval $[\bar{r}_i, \bar{r}_{i+1}]$. Its profits are given by

$$\Pi_i = \int_{\bar{r}_i}^{\bar{r}_{i+1}} \Delta(g - w_i)dr = \Delta(g - w_i)(\bar{r}_{i+1} - \bar{r}_i).$$

For a given number of firms, wages and profits at the Nash equilibrium can be determined as follows. It can be readily verified that a Nash equilibrium exists in wages. We find the Nash equilibrium wages by taking the first-order condition for Π_i with respect to w_i:

$$\frac{\partial \Pi_i}{\partial w_i} = -(\bar{r}_{i+1} - \bar{r}_i) + (g - w_i)\left(\frac{\partial \bar{r}_{i+1}}{\partial w_i} - \frac{\partial \bar{r}_i}{\partial w_i}\right) = 0. \tag{4.16}$$

Using (4.14), (4.15), and (4.16) and $r_i - r_{i-1} = L/M$, and setting equilibrium wages equal to each other, we obtain

$$w^*(M) = g - sL/M. \tag{4.17}$$

This solution is unique, for the first-order conditions are a system of linear equations in the wage of each firm. Assuming that $g > 3sL/2M$, the worker with the worst match whose training cost is $sL/2M$ will choose to work so that, in equilibrium, all workers are employed.

Assume now that there is free entry while firms remain equidistant.[13] As observed by Helsley and Strange (1990), the agglomeration of firms in the CBD has the nature of a public good (or externality) for the workers because the entry of a new firm leads to a wage increase (recall that $w^*(M)$ given by (4.17) is increasing in M).[14] This is because an additional firm improves the quality of the average match between skills and job requirements. As the number of firms keeps increasing, firms have to pay higher wages because adjacent firms compete for workers who are better matches. In the limit, when the number of firms becomes arbitrarily large, the wage approaches the competitive level g, whereas profits go to zero. However, each firm that enters the market must pay a positive fixed cost f. Therefore, at the free-entry equilibrium, the number of firms is limited.

Because profits per firm gross of entry costs are equal to $\Pi^*(M) = s\Delta L^2/M^2$, the equilibrium number of firms is

$$M^* = L\sqrt{s\Delta/f}. \tag{4.18}$$

Clearly, the long-run number of firms rises with both the thickness (Δ) and the diversity of the labor market (L) but not at the same speed. Substituting (4.18) into (4.17) yields the long-run equilibrium wage:

$$w^* = g - \sqrt{\frac{sf}{\Delta}} = g - \sqrt{sf\frac{L}{N}}.$$

Because the quality of the average match improves, the long-run equilibrium wage rises with the thickness Δ of the labor market. Thus, as expected, denser urban labor markets are associated with higher wages. In particular, since w^* is a strictly increasing and strictly concave function of N, the equilibrium wage increases at a decreasing rate when the size of the labor force rises provided that L remains constant. Stated differently, a better match gives rise to increasing returns in the aggregate. In the present setting, the city size is therefore determined endogenously according to the trade-off between increasing returns in the final sector and workers' commuting costs to the CBD. In other words, although the microeconomic mechanism vastly differs from the one investigated above, we arrive at the same conclusion when we look at what happens at the city level.

By contrast, if the population rises only through a more diversified labor force (hence, L and N grow at the same pace), the equilibrium wage is unaffected. This shows that an expansion of the labor force through a higher density or a wider range of skills may have different effects on the level of wages when workers are heterogeneous. In particular, the consequences for the city size of a growing population depend on the way it affects the parameters L and Δ of the economy.

4.3 CITY SIZE UNDER SCALE ECONOMIES

We have just seen that several mechanisms operating at the level of individual agents may generate increasing returns at the city level. Building on this observation, we assume that the city's production function exhibits increasing returns. For simplicity, here we assume that the increasing returns in production is due to the internal scale economies of a single firm that is accommodated in a city (a factory town). Our purpose is to explore, under different institutional arrangements, the implications of this assumption for the process of city formation. Specifically, we consider two types of institutions. In the first one, the city is viewed as the outcome of decisions made by a developer who internalizes the benefits of amalgamating workers within a factory. In the second, the city is formed by a community of workers who choose to combine efforts in order to enjoy the surplus generated by the presence of increasing returns.

4.3.1 The City as a Firm

Consider a monocentric city with a single firm located at the center.[15] This firm produces one good sold on the world competitive market at price p using one production factor, *labor*. The firm's production function is

$$X = F(N),$$

where X is the amount of the traded good and N the mass of workers. This function is such that $F(0) = 0$, it is strictly increasing in N, and there exists $N_a > 0$ such that

$$\frac{dF}{dN} \equiv F'(N) \gtreqless F(N)/N \quad \text{as } N \lesseqgtr N_a. \tag{4.19}$$

Thus, production involves increasing returns for $N < N_a$ and decreasing returns for $N > N_a$.

Potential employees of the firm who have the same preferences as the consumers studied in Section 3.3.1 enjoy a reservation utility level $\bar{u}$ in the rest of the economy. To be able to produce, the firm must attract some workers from the rest of the world. To this end, it must pay them a wage w high enough for these workers to be compensated for the urban rent determined on a competitive land market as well as for the commuting cost to the firm. Indeed, when some workers choose to reside within this city they anticipate that they will be organized according to a residential equilibrium such as that described in Section 3.3.1, and their income is now given by the wage w. Given this wage rate, workers will migrate into the new city as long as the utility level they can reach there is higher than or equal to their reservation utility. As seen in Section 3.3.2, the equilibrium utility level in the city decreases when the population size rises. Consequently, workers will stop migrating just when the utility level they can reach within the factory town is equal to $\bar{u}$. Clearly, such a problem belongs to the family of open city models.

When the wage rate is w, the city fringe $r^*(w, \bar{u})$ is determined by the unique solution to the equation

$$\Psi[w - T(r), \bar{u}] = R_{\mathrm{A}}. \tag{4.20}$$

Using (3.24) and (3.25), we thus obtain the equilibrium mass of workers residing in the city as given by the following function of the wage rate:

$$N(w, \bar{u}) = \int_0^{r^*(w,\bar{u})} \frac{2\pi r}{S[w - T(r), \bar{u}]} dr, \tag{4.21}$$

which is called the *population supply function* (from the rest of the economy to the city). When w increases, (4.20) implies that the bid rent curve of each worker moves upward and, hence, the urban fringe expands. Furthermore, because workers' reservation utility is fixed, their Hicksian demand for land $S[w - T(r), \bar{u}]$ falls. Consequently, (4.21) implies that the population supply rises with the wage offered by the firm. When w becomes arbitrarily large, $N(w, \bar{u})$ goes to infinity. It is also easy to show that an increase in the reservation utility $\bar{u}$ leads to a decrease in the labor supply to the city. As a result, more workers are attracted to the factory town when w increases, but less workers move in when the level of satisfaction in the rest of the economy rises.

Conversely, we may define the wage function $w(N, \bar{u})$ as the wage the firm must pay to attract exactly N workers when the reservation utility level is $\bar{u}$. Clearly, $w(N, \bar{u})$ is the inverse of $N(w, \bar{u})$. It can readily be verified that this function is strictly increasing in both N and $\bar{u}$, whereas $N \to \infty$ implies that $w \to \infty$.

Let $C(N, \bar{u})$ denote the value of the residential cost (3.26) evaluated at the efficient allocation when the population is equal to N and the utility level $\bar{u}$ with $C(0, \bar{u}) = 0$, and let

$$ADR(w, \bar{u}) = \int_0^{r^*(w,\bar{u})} [\Psi(w - T(r), \bar{u}) - R_A] 2\pi r dr \tag{4.22}$$

be the *aggregate differential land rent* (that is, the sum across locations of the difference between the urban land rent and the agricultural land rent) when the income is w and the common utility level $\bar{u}$.

It is straightforward to check that at the residential equilibrium total income equals the residential cost plus *ADR* for all values of N:

$$Nw(N, \bar{u}) = C(N, \bar{u}) + ADR(w(N, \bar{u}), \bar{u}). \tag{4.23}$$

Differentiating this expression with respect to N yields

$$w(N, \bar{u}) + N\frac{\partial w(N, \bar{u})}{\partial N} = \frac{\partial C(N, \bar{u})}{\partial N} + \frac{\partial ADR[w(N, \bar{u}), \bar{u}]}{\partial N}. \tag{4.24}$$

Because the integrand of (4.22) is zero at $r^*(w, \bar{u})$ by (3.20), we have

$$\begin{aligned}\frac{\partial ADR(w(N, \bar{u}), \bar{u})}{\partial N} &= \int_0^{r^*(w,\bar{u})} \left(\frac{\partial \Psi}{\partial w}\right)\left(\frac{\partial w(N, \bar{u})}{\partial N}\right) 2\pi r dr \\ &= \frac{\partial w(N, \bar{u})}{\partial N} \int_0^{r^*(w,\bar{u})} \frac{2\pi r}{S[w(N, \bar{u}) - T(r), \bar{u}]} dr \\ &= \frac{\partial w(N, \bar{u})}{\partial N} N\end{aligned}$$

where (3.15), (3.25) and (4.21) have been used. Therefore, (4.24) becomes

$$w(N, \bar{u}) = \frac{\partial C(N, \bar{u})}{\partial N}. \tag{4.25}$$

This means that the wage the firm has to pay to attract one additional worker is equal to the marginal residential cost. Because $\partial w / \partial N = \partial^2 C(N, \bar{u}) / \partial N^2$ and $w(N, \bar{u})$ is strictly increasing in N, we see that C is strictly convex in N. We may similarly show that C is strictly increasing in $\bar{u}$. In summary, we have

Proposition 4.1 *The residential cost is strictly increasing and strictly convex in N as well as strictly increasing in $\bar{u}$.*

Intuitively, the reason is that, because of an increase in travel distance, the total commuting costs within the city increase more than proportionally with the population size. In other words, given the monocentric structure, *there are diseconomies in urban transportation when the population rises*.[16] This result coincides with another well-documented fact in economic history that high commuting costs placed an upper limit on the growth of cities for fairly long periods (see Bairoch 1985, chap. 12). Another important implication is that the cost of living in larger cities is higher because of the higher commuting costs workers have to pay. This does not mean, however, that people residing in cities are worse off than others. As seen in Section 4.2, individuals living in larger cities earn higher wages. The increase in urban costs is the response of the market to this wage premium.

There are two equivalent ways to look at the factory town problem. Consider the first one. The firm buys some land from farmers at the price R_A and then plans and manages every aspect of the formation of the city, including the allocation of housing to workers subject to workers having the reservation utility level $\bar{u}$. The cost borne by the firm is then given by $C(N, \bar{u})$ because the residential cost is minimized at the residential allocation. In this case, the firm chooses a population size N so as to maximize its profit given by

$$\Pi(N) = pF(N) - C(N, \bar{u}).$$

If the firm were to do so, however, a tremendous amount of action would be entailed, and even more information would be required. Fortunately, there is an alternative and simpler way for the developer to maximize its profit. The firm still buys the land from farmers at R_A, but now lets the competitive market for land determine the residential allocation. In this case, the firm sets a wage w and, hence, the corresponding $N(w, \bar{u})$ workers will migrate into the city and organize themselves on the residential area, as described in Section 3.3.1. In particular, the urban land will be leased at the competitive market rent. The firm then chooses a wage w so as to maximize its profit in which the aggregate differential rent is capitalized as follows:

$$\Pi(w) = pF[N(w, \bar{u})] - wN(w, \bar{u}) + ADR(w, \bar{u}).$$

In fact, the alternatives above yield the same outcome. Indeed, the first problem

$$\max_N pF(N) - C(N, \bar{u})$$

is equivalent to

$$\max_w pF[N(w, \bar{u})] - C(N(w, \bar{u}), \bar{u}) \tag{4.26}$$

because $N(w, \bar{u})$ increases from zero to infinity when w varies from zero to infinity. In turn, by (4.23), (4.26) amounts to the second problem

$$\max_w pF[N(w, \bar{u})] - wN(w, \bar{u}) + ADR(w, \bar{u}).$$

Consider the first problem. Differentiating Π with respect to N and using (4.25) yields the following equilibrium condition:

$$pF'(N^*) = w(N^*, \bar{u}), \tag{4.27}$$

which states that the profit-maximizing employment level is such that the equilibrium wage equals the marginal value product of labor even though the firm is not a wage-taker on the labor market. For the second-order condition to be met, it is necessary that

$$pF''(N^*) - \frac{\partial^2 C(N^*, \bar{u})}{\partial N^2} < 0, \tag{4.28}$$

and thus both $F''(N^*) > 0$ and $F''(N^*) < 0$ are consistent with the second-order condition because $C(N, \bar{u})$ is strictly convex.

Figure 4.1 shows graphically that the profit-maximizing employment arises at point N^* when the tangents to $pF(N)$ and $C(N, \bar{u})$ are parallel.

Furthermore, it is easy to see that the equilibrium population size rises when the output price increases. More interesting is the impact of the transportation costs. To simplify the argument, suppose that $T(r) = tr$ and, for the city to continue to exist when t rises, that $\Pi(N^*) > 0$. Then, it is easy to check that an increase in t leads to a higher residential cost for each given value of N. Simultaneously, the wage function moves upward so that the marginal residential cost also shifts upward because of (4.25). As a consequence, the equilibrium mass of workers as well as the urban fringe shrink. Stated differently, lower

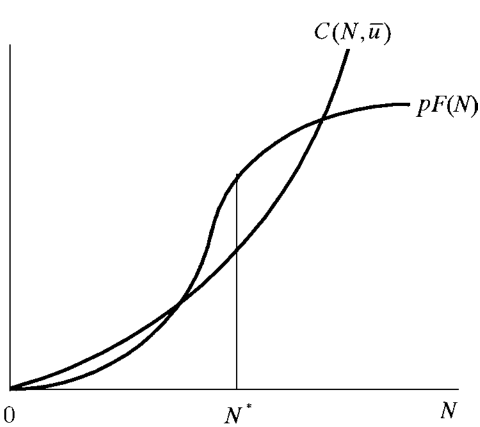


Figure 4.1: The determination of the population size in the factory town.

commuting costs lead to a larger population and a spreading out of the urban area. This is the other side of the same coin.

4.3.2 The City as a Community

Assume now that a group of workers chooses to join efforts within a community to benefit from the presence of increasing returns in production activity. For that purpose, they must work at the same place where the production function, which is now to be interpreted as the community production function, is given by $F(N)$, as described by (4.19).

These workers elect a *local government* endowed with the task of maximizing their well-being. Because all these workers are identical, they are to enjoy the same utility level, and thus the objective of the local government is to maximize this common utility level. In so doing, of course, the local government must meet some budget constraint, which will be defined below.

As usual, it is assumed that the community buys the land from farmers at R_A. To maximize the utility level of the community members, the local government chooses a certain population size N^o. Again, two approaches are possible. In the first one, the local government plans and manages all the aspects of city formation. Consequently, its budget constraint is expressed as follows:

$$pF(N) - C(N,u) \geq 0.$$

This implies that the problem of the local government is to achieve the highest utility for the community members by choosing the community size while satisfying the budget constraint, that is,

$$\max_N u \quad \text{s.t.} \quad pF(N) - C(N,u) \geq 0,$$

where N must be strictly positive for the problem to be meaningful.

Because $C(N,u)$ is strictly increasing in u, the budget constraint is binding at the optimum. Therefore, the following two conditions must hold in optimum[17]:

$$pF'(N^o) = \frac{\partial C(N^o,u)}{\partial N} \tag{4.29}$$

$$pF(N^o) = C(N^o,u). \tag{4.30}$$

The strict convexity of the residential cost function implies that

$$\frac{\partial C(N^o,u)}{\partial N} > \frac{C(N^o,u)}{N^o},$$

and thus (4.29) and (4.30) yield

$$F'(N^o) > \frac{F(N^o)}{N^o}.$$

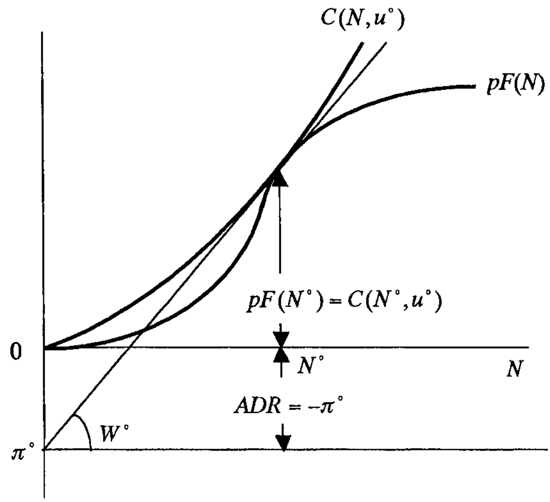


Figure 4.2: The determination of the population size in the community.

Thus, We have:

Proposition 4.2 *The utility-maximizing population size occurs in the phase of increasing returns to scale.*

Hence, *for a group of workers to agglomerate in a community and engage together in production, there must be increasing returns*; otherwise, no community would emerge. This shows how increasing returns operate to create an agglomeration of workers around the center of production. Yet, the corresponding urban area will accommodate a finite mass of workers. This is because the benefit of increasing returns will be less than the extra cost generated by an additional worker once the optimum population size is reached.

Graphically, Figure 4.2 shows that the optimum population N^o arises for the value of u^o at which $F(N)$ and $C(N, u^o)$ are tangent.

As in the foregoing, it is easy to see that u^o rises when the output price increases. If $T(r) = tr$, an increase in t leads to a decrease in u^o because the residential cost curve $C(N, u)$ moves upward with t. However, the impact of p and t on the population size is ambiguous in that it depends on the curvature of $F(N)$ at the tangency point between $pF(N)$ and $C(N, u)$.

Suppose now that the optimum population of workers decides to form a community. As usual, the residential allocation can be achieved through a competitive land market. Assume that workers are paid at their marginal value product so that their common income is

$$w^o = pF'(N^o).$$

Because the optimum population is such that $N^o w^o > pF(N^o)$, the wage bill exceeds the output value, thus implying that the production activity generates a loss. Nevertheless, rewriting (4.23) at the optimum yields the following relationship:

$$ADR(N^o, u^o) = -[pF(N^o) - N^o w^o]. \tag{4.31}$$

Accordingly, at the optimum the aggregate differential land rent capitalized by the community just compensates for the loss incurred in the production activity once workers receive the marginal value product of labor. This is an important result suggesting how the loss incurred in the production under increasing returns may be compensated through the capitalization of the differential land rent, without introducing any distortive tax. This is a version of a more general relationship named the *Henry George theorem* by Stiglitz (1977).[18]

We are now able to compare the two institutional systems studied in the foregoing sections. It is straightforward from (4.27) through (4.30) that the factory town and the community models yield the same outcome if and only if firms' profits are zero. Stated differently, the profit-maximizing size equals the optimum size if and only if the reservation utility level is the same as the optimum utility level.

This is no longer true when firms' profits are positive. The two outcomes differ in the following way: the profit-maximizing firm chooses a labor force strictly larger than the optimum size when the reservation utility level $\bar{u}$ is lower than the optimum utility level u^o. Indeed, if $\bar{u} < u^o$, the firm is able to pay its workers a wage high enough to attract more than N^o workers. This induces the firm to hire a population larger than the optimum one. On the other hand, when $\bar{u} > u^o$ the firm cannot break even and, hence, the factory town is not formed.

4.3.3 System of Cities: A Simple Framework

The foregoing results suggest an interesting entry – exit process leading to the formation of a *system of cities*. To this end, suppose the economy consists of an urban sector with many potential cities and a rural sector with an insignificant population (such as in most developed countries). For simplicity, we assume that the total population of the urban sector is exogenously given by a constant $\mathcal{N}$. Because the economy under consideration is closely tied with the world economy, both the output price of the urban sector and that of the rural sector are assumed to be given and fixed. Whenever a city is to be developed, the production function is the same and is expressed by (4.19). The economy has enough land for each developer to buy the land needed at the fixed agricultural land rent R_A and for all the created cities not to overlap.

The process of city formation works as follows. We have seen that u^o is the highest utility a firm-developer may sustain without losses, and w^o is the highest wage it can pay to its workers. Suppose that the optimum number of cities in the economy, given by $\mathcal{N}/N^o$, is large enough to be safely treated as a real number.[19] Because the size of each city is small compared with the population $\mathcal{N}$, each developer will consider the utility level in the other cities as given. As long as $u^* < u^o$, each existing developer can earn a strictly positive profit, for example by hiring N^o workers at a wage lower than w^o, thus inviting the entry of new developers. Eventually, profits earned from the formation of new cities will become zero and entry will stop. At this zero-profit equilibrium $u^* = u^o$, and each firm will hire exactly N^o workers and will pay them w^o. This implies that workers enjoy the optimum utility level u^o in each city and that the number of cities developed will be optimal.

Clearly, the same results hold if competition takes place among utility-maximizing communities. Furthermore, because all cities are identical, (4.31) holds in each one of them. We may summarize these results as follows:

Proposition 4.3

1. *The equilibrium city system resulting from competition between profit-maximizing firms or developers is identical to the optimum system in which the common utility level is maximized, and conversely.*
2. *In every city of the equilibrium – optimum city system, the aggregate differential rent just covers the loss from production activity when goods and labor are priced at their competitive level (Henry George theorem).*

It follows from the first part of the proposition that the optimum city system may be obtained through competition among profit-maximizing firms or utility-maximizing communities.[20] This is interesting because, for factory towns to exist in the zero-profit equilibrium, the firms' production function must exhibit increasing returns. Although workers are paid at their marginal productivity, which exceeds the average productivity, the corresponding loss incurred in production is exactly compensated by the aggregate differential land rent capitalized within each factory town. This result, suggested by Vickrey (1977), shows how the land market may provide a solution to a well-known problem in economic theory due to increasing returns, which is usually solved by adding constraints that lead to second-best solutions.

Furthermore, in equilibrium all individuals enjoy the same utility level regardless of the city in which she resides. This implies that the corresponding allocation is *fair* in the sense of Rawls. This is an intrinsic property of the spatial equilibrium that suggests that a market economy is not necessarily characterized by spatial inequalities, as is often claimed by the proponents of regional planning. Within the present simple setting, there is no reason for the government to interfere with the market system when such an intervention is motivated by redistributive considerations because the market outcome maximizes the Rawlsian welfare function.

Finally, because the optimum population changes with the price p, the equilibrium urban system of an open economy is affected by the external factors ruling the world market.

However, all cities are identical and do not trade among themselves. The city system obtained here looks like a collection of isolated cities, whereas one would expect to get cities of different types trading goods. We will return to this matter below.

4.3.4 Notes on the Literature

The work of Mills (1967) has served as a basis for many developments in urban economics regarding especially the role of developers and the creation of communities. In this respect, our approach to the formation of a system of cities owes much to Mirrlees (1972), Henderson (1974), Eaton and Lipsey (1977), Kanemoto (1980, chap. 2), Berglas and Pines (1981), Hochman (1981), and Fujita (1989, chap. 5). Schweizer (1986) has provided a discrete treatment of city numbers. The Henry George theorem, as applied to the financing of the loss incurred under perfect competition and increasing returns, corresponds to an old conjecture stated by Hotelling in a series of papers published in the late 1930s, which are available in Darnell (1990). This theorem has been studied independently by Serk-Hanssen (1969) and Starrett (1974), whereas Vickrey (1977) did the same but within a system of competing cities.[21] Finally, Papageorgiou and Pines (2000) have shown that the optimality of the market outcome obtained in Proposition 4.3 depends critically on the fact that cities are many and replicable.

4.4 TRADE IN A SYSTEM OF CITIES

4.4.1 Specialization and Trade

The urban system we have just described looks like a system of "isolated city-states" (Papageorgiou). This is a major limitation because exchange and trade are often considered the main ingredients of urban systems. However, the preceding results can be extended to the case of several goods that are traded between any two cities. In this section, we briefly describe the general equilibrium model developed by Henderson (1974; 1987; 1988, chap. 2); the reader is referred to these works of Henderson as well as to Becker and Henderson (2000) for further details.[22]

The main question to be solved is, Why is it that cities trade goods instead of each city producing all of them? We know from the spatial impossibility theorem that some market failures are necessary for trade to arise. In this section, agglomeration and trade are not inconsistent because scale economies are external to firms but internal to the industry, as in A. Marshall, thus, firms producing the same good benefit from being located together (see also Chapter 8). On the other hand, we have seen that adding a new worker to the city generates a higher average per person commuting cost (see Proposition 4.1). As a result, when external economies are confined within each industry, specialization will lead to a better exploitation of scale effects. Indeed, for the same degree of global efficiency in the two sectors, when workers residing within a city are engaged in the production of two different tradable goods, the average commuting cost is higher than when they are located in two separate cities producing each tradable good. Consequently, when the transport costs of these goods are sufficiently low (here they are assumed to be zero), it is more efficient for each city to supply a single tradable good that is both locally consumed and exported to the rest of the economy, whereas other tradable items are imported from other cities.

But why should such an efficient outcome occur in a market economy? According to Henderson (1974), the answer is to be found in the existence of some economic agents (the "city corporations") who understand that they may benefit from organizing cities in a way that maximizes the utility level of residents while internalizing the external effects generated by the agglomeration of firms and workers belonging to the same industry.[23] As discussed above, examples of such agents include land developers and local governments. These agents are assumed to be many, whereas the total population of workers is sufficiently large for many cities of each type to exist. In order to see how Henderson's model works, we present here a simplified version in which labor is the only production factor.

Consider a representative city. For the reason discussed in the preceding paragraph, it is never profitable for a developer to produce more than one traded good in the city she develops, and thus a city may be identified with the traded good i produced there.[24] This means that an i-type city produces a traded good X_i in addition to housing H_i.[25]

Technology in the i-sector exhibits constant returns to scale at the firm level. However, there are external economies at the industry level such that the production function of this sector is as follows:

$$X_i = E_i(N_i)N_{if},$$

where N_{if} is the number of residents working in the i-sector, whereas N_i is the number of city residents and $E_i(N_i)$ a Hicksian shift factor taken by each firm as given, with $E_i' > 0$.

In each city, housing is produced under constant returns to scale at the firm level. Instead of introducing the details of commuting and land consumption, it is simply assumed that a negative externality is at work in the housing sector:

$$H_i = E_h(N_i)N_{ih},$$

where N_{ih} is the number of residents working in this sector, whereas $E_h(N_i)$, with $E_h' < 0$, stands for the external diseconomies in the housing sector when the total population of residents is N_i. Because the two "scale effects" are external to firms, all markets may be assumed to be perfectly competitive and firms to be price-takers (Chipman 1970).

Henderson (1987) assumes particular functional forms for these two external effects:

$$E_i(N_i) = \exp(-\varphi_i/N_i),$$

where $\varphi_i > 0$ is a measure of the degree of increasing returns in the production of good i, and

$$E_h(N_i) = N_i^{-\delta},$$

where $\delta > 0$ expresses the intensity of commuting and congestion costs within a city.

There are n traded goods, and consumers have the same Cobb–Douglas preferences given by

$$U = x_1^{a_1} \cdots x_n^{a_n} h^b,$$

where x_i stands for the consumption of good i produced in city i and h is the housing consumption. Consumers live in the city where they work and spend their income.

The role of a developer is to choose a tradable good to be produced while letting competitive markets determine individual production activities in the chosen sector as well as in housing. Because consumers are identical and free to move, they must reach the same equilibrium level regardless of the city in which they live. As in the preceding section, the equilibrium size of each type i-city is such that the common utility level of the residents is maximized at this equilibrium size; otherwise, there would exist potential profits to be earned. Stated differently, through competition in the market for cities, the benefits of agglomeration are eventually transferred to workers in order to make up for the higher urban costs they bear in more productive (and larger) cities. In this process, each developer is assumed to be able to choose the population size through setting the urban wage while taking the utility level as given (as in Section 4.2.1) or by controlling land development and usage.

Given some large population of workers in the whole economy, it is standard to determine consumers' demands for each tradable good and housing as well as the corresponding equilibrium competitive prices. Following Henderson (1974), it can then be shown that the equilibrium population size in a type i-city is such that

$$N_i^* = \frac{a_1 + \cdots + a_n}{b\delta}\varphi_i.$$

Hence, the equilibrium population of a city producing good i increases with the degree of increasing returns (φ_i) and the intensity of preferences for all tradable goods ($a_1/b + \cdots + a_n/b$), but decreases with commuting and land use costs (δ) in the urban arrangement.

Because the degree of increasing returns varies with the good produced, it is clear that cities specializing in the production of different goods have different sizes. Larger cities are those specialized in the production of traded goods with higher degrees of increasing returns. In other words,

> City sizes vary because cities of different types specialize in the production of different traded goods, exported by cities to other cities or economies. If these goods involve different degrees of scale economies, cities will be of different sizes because they can support different levels of commuting and congestion costs. (Henderson 1974, 640).

In addition, all individuals enjoy the same utility level in equilibrium, whatever the type of city in which they live, since they are free to choose where to reside. However, because the production functions are not necessarily the same, the wages paid in cities of different types are not the same, thus generating different population sizes, which in turn implies different commuting costs and land rents. Specifically, wages (w_i^*) and housing prices (R_i^*) rise with the city size in order to keep the utility level identical across cities:

$$\frac{w_i^*}{w_k^*} = \frac{R_i^*}{R_k^*} = \left(\frac{N_i^*}{N_k^*}\right)^{1/\alpha} = \left(\frac{\varphi_i}{\varphi_k}\right)^{1/\alpha}.$$

Hence, the wage ratio and the housing price ratio between two types of cities both equally rise with the ratio between their respective degrees of increasing returns. Clearly, differences in wages do not correspond here to differences in workers' well-being.

Finally, the ratio of numbers of each type of city is obtained by equating demand and supply for each type of good within the whole economy. If m_i^* denotes the equilibrium number of cities of type i, then for any two types of cities i and k it can be shown that

$$\frac{m_i^*}{m_k^*} = \frac{a_i}{a_k}\left(\frac{N_i^*}{N_k^*}\right)^{-1-1/\alpha} = \frac{a_i}{a_k}\left(\frac{\varphi_k}{\varphi_i}\right)^{1+1/\alpha}.$$

Hence, the trade-off between increasing returns and commuting costs within cities takes the following particular form:

1. As increasing returns get stronger in the production of good i, the relative number of cities of type i decreases, whereas these cities become larger.
2. When commuting costs rise, the relative number of type i-cities increases (decreases) when $\varphi_i < \varphi_k$ ($\varphi_i > \varphi_k$), whereas each city gets smaller. The number of cities of each type also depends on the expenditure share on each tradable good.

This model explains the existence of an urban system with different types of cities as well as intercity trade involving different goods. It also sheds additional light on the trade-off between increasing returns and commuting costs. However, the model does not permit one to predict the location of cities nor does it explain the urban hierarchical structure. A full theory of urban systems should explicitly account for city location. This question is investigated in Chapter 10.

4.4.2 Diversified versus Specialized Cities

Another critical issue is why some cities are *specialized* in the production of a few goods and services, whereas others are more *diversified* (Abdel-Rahman 2000; Duranton and Puga 2000). From the historical viewpoint, it should be clear that cities were diversified when intercity trade was very costly. Indeed, despite increasing returns, it was desirable to provide a wide array of goods locally to save on transport costs (Abdel-Rahman 1996). Nowadays, intercity trade costs are lower, and it seems desirable to take advantage of increasing returns. This means that cities should specialize (as discussed in Section 4.4.1). Yet, we observe that large and diversified metropolises keep growing.

One reason for diversification is the possibility of exploiting the scale economies associated with a large variety of common intermediate goods and public services (Goldstein and Gronberg 1984). If the demand for the final output of a city-industry is not perfectly elastic, then it is not profitable for this industry to grow beyond some limit. If two different industries colocate within the same city, they may enjoy a larger intermediate sector and more public services, making each industry more productive. On the other hand, this colocation requires more workers within the same city and, therefore, longer commuting. As a consequence, firms in the two final sectors must pay a higher wage. In this case, the balance is between the productivity gains of the two industries resulting from a larger intermediate sector and the commuting cost within the diversified city. The same type of argument may be applied to consumption (Abdel-Rahman 2000).

A second reason is that the diversified city is able to smooth out random shocks affecting specific urban industries. In this case, the diversified city is to be viewed as a portfolio of activities. When one activity is adversely affected, workers have the opportunity to move to other sectors. The expected wage is then higher than in a specialized city (Krugman 1991b).

Finally, Jacobs (1969) has argued that urban diversity facilitates innovation because it allows new producers to observe and borrow ideas initiated by firms belonging to other industrial sectors. Duranton and Puga (2002) have shown how uncertainty about the production process in the initial phases of the product cycle may lead to the coexistence of different sectors within the same city. When firms master their production processes, they relocate to more specialized areas, and thus the diversified cities are viewed here as "nurseries."

4.5 COMPETITION AND THE SPATIAL ORGANIZATION OF MARKETS

We now come to quite a different tradition, which is deeply rooted in classical location theory. If firms and consumers are geographically dispersed and the number of firms is small relative to the mass of consumers owing to indivisibilities in production, each firm has some monopoly power over the consumers in its immediate vicinity. In other words, the presence of increasing returns at the plant level prevents spatial markets from being perfectly competitive because differences in consumer locations, and hence transport costs, are a source of market power. Competition in space is, therefore, imperfect and should be studied according to the relevant theories. Once this is recognized, the trade-off between increasing returns and transport costs turns out to be crucial for determining the number of firms competing within a given area whose population size is given. The present setting may be viewed as the "primal" of that studied in Section 4.2.2.

4.5.1 Equilibrium and the Number of Firms in Space

The prototype of spatial competition is generally attributed to Hotelling (1929), who studied firms' price decisions assuming that consumers' locations are fixed.[26] Hotelling's main purpose was to model competition so that each firm's demand is continuous, while permitting consumers to react discontinuously at the individual level.[27] The heterogeneity across consumers, introduced through transportation costs, ensures that individual discontinuities stemming from consumers' mutually exclusive purchases are distributed so as to be unnoticeable to the firm.

Yet, the essence of spatial competition was probably better described by Kaldor (1935). According to this author, locations in space mold the nature of competition between firms in a very specific way. Whatever the number of firms participating in the aggregate, *competition is localized: each firm competes more vigorously with its immediate neighbors than with more distant neighbors.* Spatial competition is therefore inherently strategic in that each firm is only concerned with a small number of direct competitors regardless of the total number of firms in the industry. This does not imply, however, that the industry is formed by independent clusters of sellers. Because a chain connects any two firms such that any two subsequent firms in the chain are direct competitors, all of them are interrelated within a complex network of interactions. Some "chain effect" linking of apparently independent firms seems to be inherent to the spatial framework. The whole demand system must then be inspected to delineate the extent of a spatial market.

To gain some insights into the working of competition in such a context, we consider a very simple setting in which M firms supplying the same good are distributed equidistantly along a circle C of length l while consumers are continuously distributed along the same circle.[28] Then, firm i has two *direct competitors*, firms $i - 1$ and $i + 1$. The market situated between firms $i - 1$ and $i + 1$ is segmented according to the principle stated above: each consumer patronizes the firm with the lowest full price. Hence, for a vector of prices, there are three groups of consumers in this local market: the customers of firm $i - 1$, of firm $i+1$, and of firm i. A unilateral price cut by firm i will consequently extend its own market only at the expense of firms $i-1$ and $i+1$, whereas the other firms are not directly affected. Therefore, the cross-price elasticity between firm i and firm $j \neq i-1, i+1$ is zero.

Each consumer purchases one unit from the firm that, for him, has the lowest full price, which is defined as the posted price plus the transportation cost to the corresponding firm. The consumers are thus divided into different segments, and each firm's demand is the sum of consumers' demands in one particular segment. The boundary between two firms' markets is given by the location of the consumer indifferent between them known as the *marginal consumer*. This boundary is endogenous because it depends on the prices set by the firms. Given the continuous dispersion of consumers, a marginal variation in price changes the boundary and each firm's demand by the same order.

To keep things simple, we assume that the density of consumers over the circle C is uniform (maybe because the lot size is fixed and the same across locations) and equal to $n > 0$. Each consumer buys one unit of the product from one of the M firms located at equal distance l/M.[29] A consumer at location $r \in C$ has an indirect utility given by

$$V_i(r) = u + Y - p_i - t|r - r_i| \tag{4.32}$$

when she patronizes the shop located at $r_i \in C$. In this expression, u describes the gross utility that a consumer derives from the product, Y her income, p_i the price posted by firm i, $t|r - r_i|$ the transportation cost the consumer must bear when visiting firm i with $t > 0$, and $|r - r_i|$ the length of the shortest arc connecting r and r_i. The net income $Y - t|r - r_i|$ is supposed to be high enough for each consumer to be able to buy the good. All firms have the same production cost given by $C(q) = f + cq$, where $f > 0$ and $c > 0$ (see also (4.2)). A price vector is denoted by $\boldsymbol{p} = (p_1, \ldots, p_M)$, whereas $\boldsymbol{p}_{-i}$ stands for the price vector $\boldsymbol{p}$ from which the ith component has been deleted.

Our first task is to determine the position of the marginal consumer between firm i and each of its two neighbors. Consider the case of firm $i-1$. The corresponding marginal consumer is located at $r_{i-1,i} \in [r_{i-1}, r_i]$, which must satisfy $V_{i-1}(r_{i-1,i}) = V_i(r_{i-1,i})$ so that

$$r_{i-1,i} = \frac{p_{i-1} - p_i + tl/M}{2t}.$$

Because a similar expression holds for $r_{i,i+1}$, it follows that the demand to firm i when prices are given by p_{i-1}, p_i and p_{i+1} is

$$D_i(p_{i-1}, p_i, p_{i+1}) = n\frac{p_{i-1} - 2p_i + p_{i+1} + 2tl/M}{2t}. \tag{4.33}$$

This expression reflects the localized nature of competition, for, besides its own price, D_i depends only upon the prices charged by its two neighbors.

Consequently, firm i's profit, contingent on the prices p_{i-1} and p_{i+1} set by its two neighbors, can be written as follows:

$$\pi_i(\boldsymbol{p}) = (p_i - c)D_i(p_{i-1}, p_i, p_{i+1}) - f. \tag{4.34}$$

We consider a Nash price equilibrium in pure strategies, that is, a price vector $\boldsymbol{p}^*$ such that each firm $i = 1, \ldots, M$, anticipating correctly the prices charged by the other firms, maximizes its profit $\Pi_i(p_1^*, \ldots, p_i, \ldots, p_M^*)$ at p_i^*. If such an equilibrium exists, it must solve the first-order condition applied to (4.34):

$$p_{i-1} - 4p_i + p_{i+1} + 2tl/M + 2c = 0 \quad i = 1, \ldots, M.$$

This is a system of M linear equations that has a unique solution given by

$$p^* = p_i^* = c + tl/M. \tag{4.35}$$

This solution is a Nash equilibrium. Indeed, the second-order condition is locally satisfied, whereas any unilateral deviation that prices firms $i - 1$ and $i + 1$ out of business leads to negative profit.[30]

Inspecting (4.35) reveals that firms apply an absolute markup tl/M that increases with the transportation rate t as well as with the distance l/M between two adjacent firms. In other words, geographical isolation, economically expressed by the value of the transportation cost, allows each firm to have market power. However, this market power is restricted by the market power exercised by the closest competitors, as measured by the distance between two successive firms along C. This shows how space acts as a barrier to competition: higher transportation costs, fewer firms, or both yield higher equilibrium price and profit.

On the contrary, when the number of firms becomes arbitrarily large, the equilibrium price converges toward the marginal production cost, that is, the competitive outcome. But the existence of a fixed cost will prevent the number of firms from rising indefinitely. In fact, firms are confronted by a trade-off when deciding whether or not to enter. A firm will only enter if it can locate sufficiently far from other firms (in terms of economic distance, not physical distance) so that it can serve enough consumers and charge a high enough price to cover its fixed costs.

The equilibrium number of firms M^* is obtained when firms' equilibrium profit is equal to zero (disregarding again the integer problem), that is, when $(tl/M)(nl/M) - f = 0$ so that

$$M^* = l\sqrt{\frac{nt}{f}}. \tag{4.36}$$

Consequently, *the equilibrium number of firms increases with the unit transportation cost but decreases with the fixed production cost.* The intuition is the same as presented in Section 4.2.3. This is the spatial competition version of the now classical trade-off between increasing returns and transportation costs:

Whereas the former reduces the average production cost, it also increases the cost of transportation for those traveling to the firm. This trade-off determines the number of firms in space.

This result can be given an interesting interpretation from the historical point of view. When technologies were inefficient, production involved low investments, and transport was very costly, a large number of firms were operating at small scales. Since the beginning of the Industrial Revolution, however, transport costs have declined dramatically whereas production has entailed increasing overhead costs. Consequently, we may safely conclude that *the type of technological progress observed for many decades in developed countries has led to a substantial reduction in the number of operating plants as well as to an expansion of their size and market area.* On the other hand, a rise in the population size through an increase of n or a geographical expansion of the economy expressed via an increase of l both lead to a larger number of plants (Cain 1997).

4.5.2 The Optimum Spatial Distribution of Firms

Consider now the efficient configuration. How many firms should there be in the market? Or, equivalently, how many markets should be open? In the setting considered in the section above in which each consumer buys one unit of the product, there is no deadweight loss associated with a discrepancy between price and marginal cost. Because consumers' indirect utility (4.32) is linear in income and each consumer buys one unit of the good produced at a constant unit cost c, the optimal number of firms minimizes the firms' fixed production costs plus consumers' total transport costs. Hence, the problem involves a trade-off, for increasing the number of firms, hence fixed costs, reduces the aggregate transport costs, and vice versa.[31]

It can readily be verified that the social cost to be minimized is defined as follows:

$$\begin{aligned} C(M) &= Mf + 2M \int_0^{l/2M} ntr dr \\ &= Mf + ntl^2/4M. \end{aligned}$$

Treating M as a real number and differentiating this expression with respect to M yields the optimum number of firms (the second-order condition is met because C is strictly convex):

$$M^o = \frac{l}{2}\sqrt{\frac{nt}{f}} = M^*/2. \tag{4.37}$$

Using (4.36), we thus have shown the following:

Proposition 4.4 *In the spatial competition model, the equilibrium number of firms is larger than the optimal number.*

This is a fairly general result suggesting that the market tends to provide too many small firms, thus leading to a denser pattern of production than the one chosen by a benevolent and informed planner.[32] The reason is that, although there are M firms in the industry, each firm competes with its two neighbors only, thus leading to high market prices, which, in turn, invite more entry. Proposition 4.4 can be viewed as the spatial counterpart of Chamberlin's excess capacity theorem in that the total capacity built by the market, expressed by the number of plants, is too high. This is interesting because proponents of regional planning often argue that the market works poorly in fighting against regional imbalances and "desertification." After all, the market seems to generate denser patterns of production than what is optimum – a result that invites us to be careful in evaluating the relevance of some criticisms of a spatial market economy's organization.

4.5.3 Land Capitalization

Space has an important consequence that is ignored in spatial competition models but is central to urban economics, namely *capitalization*. Capitalization means here that the land rent reflects both the price paid and the transport cost incurred by the occupant when patronizing the cheapest store. There is no capitalization in the standard spatial competition model because land is not involved. Yet, the model can be extended to cope explicitly with consumers simultaneously choosing location and consumption. The basic framework has been laid down by Fujita and Thisse (1986) as well as by Asami, Fujita, and Thisse (1993); it is described in the paragraphs that follow.

We suppose that firms and households make their decisions sequentially. In the first stage, M equidistant firms choose prices in a noncooperative way. In the second stage, given the decisions made by firms, households consume one unit of land (the lot size is fixed) in addition to one unit of firms' output. Hence, households have to choose a location and the firm to patronize.[33] For simplicity, we assume that the mass N of consumers is equal to l so that consumer density must be one ($n = 1$). Consumers compete for land and pay the land rent. Besides land rent, the income of each household is spent on the firms' output, transport cost, and the numéraire. A consumer at location $r \in C$ now has an indirect utility given by

$$V_i(r) = u + Y - p_i - t|r - r_i| - R(r) \tag{4.38}$$

when she patronizes the shop located at $r_i \in C$ and pays the market land rent $R(r)$ at location r.

When choosing their prices firms anticipate consumers' responses, thus reflecting the fact that the firms have more market power than consumers. Our equilibrium concept can then be summarized as follows. Given a price configuration of firms, consumers decide on their location and shopping place at the corresponding residential equilibrium, which is of the competitive type. With respect to firms, consumers could be viewed as the followers of a Stackelberg game in which firms would be the leaders. Firms choose prices at the Nash equilibrium of a noncooperative game whose players are the firms and in this way anticipate consumers' residential equilibrium.

To show how spatial competition may yield the first best outcome, we propose an institutional system in which the firms' payoff functions are modified through land capitalization. Specifically, we assume that firm i's payoff is defined by the sum of the operating profit made from selling its product and the added land value as a result of its presence in the urban area. The added land value ($\Delta_i(\boldsymbol{p})$) is obtained by subtracting the aggregate land rent when firm i does not operate ($ADR(\boldsymbol{p}_{-i})$) from the aggregate land rent when firm i operates with the other $n-1$ firms ($ADR(\boldsymbol{p})$):

$$\Delta_i(\boldsymbol{p}) \equiv ADR(\boldsymbol{p}) - ADR(\boldsymbol{p}_{-i}),$$

and thus firm i's payoff is

$$\Pi_i(\boldsymbol{p}) = \pi_i(\boldsymbol{p}) + \Delta_i(\boldsymbol{p}).$$

Given what we said in Section 4.3, this means that firm i behaves like a land developer within its market area, which corresponds to a segment of the urban area.

If M equidistant firms have entered the urban market and incurred the fixed cost f, each of them is active in the price equilibrium. At the resulting residential equilibrium, all consumers must reach the same utility level, and thus (4.38) implies a constant $\bar{R}$ that equals the common urban cost borne by consumers,

$$R(r) + \min_{j=1,\dots,M}\{p_j + t|r - r_j|\} = \bar{R}, \tag{4.39}$$

for all $r \in C$. In what follows, it is supposed that all firms consider the common urban cost $\bar{R}$ as a given constant.[34]

When firm i is the only nonactive firm, we have

$$R_{-i}(r) + \min_{j\neq i}\left\{p_j + t|r - r_j|\right\} = \bar{R},$$

where $R_{-i}(r)$ is the land rent prevailing at r when firm i is inactive. Then the added land value of firm i, $\Delta_i(\boldsymbol{p})$, is given by the shaded area in Figure 4.3, which can be shown to be equal to

$$\Delta_i(\boldsymbol{p}) = \frac{(p_{i-1} - p_i + tl/M)(p_{i+1} - p_i + tl/M)}{2t},$$

and thus firm i's payoff is given by

$$\Pi_i(\boldsymbol{p}) = (p_i - c)D_i(p_{i-1}, p_i, p_{i+1}) + \Delta_i(\boldsymbol{p}), \tag{4.40}$$

where $n = 1$ in (4.33).

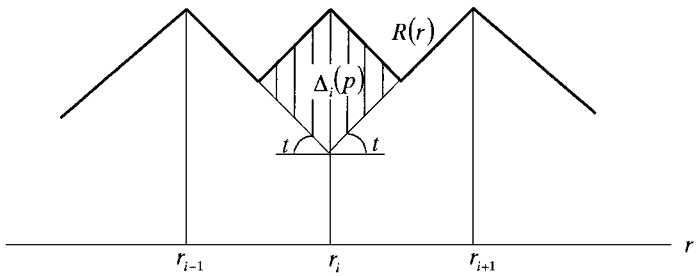


Figure 4.3: Firm i's added land value.

Differentiating (4.40) with respect to p_i and equating the resulting expression to zero, we obtain

$$\begin{aligned}\frac{\partial \Pi_i}{\partial p_i} &= D_i + (p_i - c)\frac{\partial D_i}{\partial p_i} + \frac{\partial \Delta_i(\mathbf{p})}{\partial p_i} \\ &= D_i + (p_i - c)\frac{\partial D_i}{\partial p_i} - D_i \\ &= -\frac{p_i - c}{t} = 0,\end{aligned}$$

which implies that $p_i^* = c$. Clearly, $\Pi_i(\mathbf{p})$ is concave in p_i so that the unique solution to the first-order condition is the unique price equilibrium. In fact, because $p_i^* = c$ is the optimal strategy for firm i *regardless* of the prices charged by all other firms, we have the following result:

Proposition 4.5 *Assume that firms capitalize the added land value on their market area into their payoffs. If firms consider the common urban cost as a constant, then marginal cost pricing is the equilibrium in strongly dominant strategies of the price game.*

The intuition behind this somewhat surprising result is that land capitalization is equivalent to first-degree price discrimination, up to a constant that is the same across firms because the consumer surplus is identical to the land rent (up to the same constant for all firms), which is now captured by the agent who creates it. Therefore, making the socially optimal decision is a Nash equilibrium (Spence 1976). We have, however, a stronger result because our equilibrium is in dominant strategies.

Marginal cost pricing implies that operating profits π_i are zero for all firms so that their payoff is equal to their added land value. In the entry process, the number of firms will be such that the added land value of the last firm is equal to its fixed entry cost (neglecting the integer problem). Because the added land value by a firm is equal to the corresponding reduction in total transport costs, the free-entry condition means that the last comer must balance the social gain and the social entry cost. Hence, we have the following:

Proposition 4.6 *Assume that firms capitalize the added land value on their market area into their payoffs. If firms consider the common urban cost as a constant, then the free-entry equilibrium leads to the socially optimal number of firms.*

The equilibrium number of firms is now given by (4.37). Indeed, Proposition 4.5 implies that the social value of an additional firm is identical to its added land value in the present model. Hence, when land rents are accurately accounted for, there is no excess capacity in a spatial economy, thus "proving" the following conjecture stated by Hotelling (1938, 242) more than 70 years ago:

> All taxes on commodities, including sales taxes, are more objectionable than taxes on incomes, inheritances, and the site value of land; and . . . the latter might be applied to cover the fixed costs of electric power plants, waterworks, railroads, and other industries in which the fixed costs are large, so as to reduce to the level of marginal cost the prices charged for the services and products of these industries.

Proposition 4.6 also means that we have obtained a Henry George rule in an environment in which firms behave strategically. In other words, allowing firms to capitalize the added land rent they create by their activities leads to the first best outcome even though firms behave strategically on the product market. On the other hand, they are not supposed to be able to manipulate the urban cost borne by the consumers.[35] This behavioral assumption is very much in the spirit of Hart (1985), who has suggested that firms should account only for some effects of their policy on the economy as a whole. This seems reasonable because such computations are likely to be beyond their reach. The utility-taking assumption is also acceptable as long as no firm is dominant within the economy – a claim that opens new perspectives on antitrust policy.[36]

4.5.4 Notes on the Literature

Since the path-breaking article of Hotelling (1929), it has been well known that the framework used to describe a spatial industry permits one to study the working of a differentiated industry. The power of this analogy has been exploited extensively in industrial organization (Eaton and Lipsey 1977; Gabszewicz and Thisse 1986). Beckmann (1972b) and Stern (1972) were the first who formalized in a precise way the trade-off between increasing returns and transportation costs under spatial competition, whereas the circular model was developed independently and later on by Salop (1979). There is a substantial literature on industrial organization that addresses the issue of whether the market under- or overprovides differentiated products that is of direct relevance for spatial competition. Surveys can be found in Eaton and Lipsey (1997), Gabszewicz and Thisse (1992), and Anderson, de Palma, and Thisse (1992, chap. 6). Launhardt's contribution to spatial competition was ignored until recently outside the German-speaking community. A modern presentation of his model can be found in Dos Santos Ferreira and Thisse (1996). Finally, it is worth noting that many results obtained in spatial competition were anticipated by Vickrey in his *Microstatics* published in 1964. The main extracts have been reproduced in Anderson and Braid (1999).

4.6 CONCLUDING REMARKS

We have seen that the trade-off between increasing returns and transportation costs may take very different forms and explain a wide range of issues in a spatial economy. First, this trade-off lies at the source of the mechanism governing the size of a city in either context of the city as a firm or a community. The approach taken in these two modeling strategies has much to share with club theory, as developed by Buchanan (1965) and Berglas (1976).[37] Not surprisingly, therefore, free entry in the development industry leads to the emergence of the optimum city system.[38] Although the description of a city system provided here is very simple, it is interesting that the first best outcome is achieved in a setting with increasing returns and perfect competition. This is due to the Henry George rule, which states how the loss incurred in production may be financed by the differential land rent once all goods are sold on competitive markets. It is worth noting that there is no difference in models of imperfect competition of the product or factor market provided that the differential land rent is accurately redistributed among firms. When several goods can be traded at no cost, most of these results remain valid, each city being specialized in the production of one tradable good.

Second, a similar trade-off arises in the study of the spatial organization of an industry. High overhead costs in production lead to a sparse distribution of firms, whereas high transportation costs have the opposite effects. These are the elements considered (implicitly) by Christaller ([1933] 1960) and (explicitly) by Lösch ([1940] 1954) in their explanation of a one-commodity spatial market. Increasing returns prevent spatial markets from being complete, and, as seen in the preceding sections, the number of active firms depends on the trade-off between increasing returns and transportation costs. In general, the two systems lead to a different number of markets. However, when firms are allowed to capitalize the land rent they create over their market area, the market once again yields the first best outcome. This result shows the power of the urban land rent in a market economy and reconciles Hotelling (1929) with Hotelling (1938) in addition to being a superb illustration of the contributions made by this author to the development of modern economics through the use of space as a specific economic category.

Observe that the forces at work are not tied to the particular spatial price policy used in the foregoing developments, which is called *mill pricing*. For example, the ability to engage in discriminatory pricing (see Chapter 7 for more details) implies that a firm can cut its prices in one part of its market without the need to change its prices elsewhere. This effectively reduces the power of the firm to commit to a set of prices and so strengthens price competition. But fiercer price competition will not necessarily work to the benefit of consumers. On the one hand, for a given number of firms, stronger price competition will indeed reduce the prices charged by incumbent firms. On the other hand, stronger price competition can also be expected to act as an entry deterrent and so may benefit incumbent firms. Norman and Thisse (1996) have shown that the entry deterrence effect is dominant: discriminatory pricing leads to less and larger firms than mill pricing. The power of price discrimination to limit the opening of new marketplaces in the spatial economy is particularly marked when relocation costs are high. To the extent that regional planning is concerned with encouraging greater geographical dispersion of industries, mill pricing might turn out to be more socially desirable in terms of consumer welfare and regional planning.

More fundamental, perhaps, the same trade-off is encountered in planning the number and location of all sorts of private or public facilities such as schools, recreational facilities, fire stations, and the like. The corresponding optimization problem, known as the simple plant location problem, is at the heart of facility location analysis in operations research (see Section 2.5.2). Last, the same trade-off has been once more "rediscovered" in the new economic geography that started with the work of Krugman (1991a, b). It operates in a still different way and will be studied in Part III. So it does not seem an exaggeration to say that *the trade-off between increasing returns and transportation cost is fundamental for the operation of a spatial economy.*

Two comments are still in order. First, in this chapter, we have focused on what are probably the two main existing spatial models. The first one (Sections 4.3 and 4.4) involves land and consumer location, but it does not deal with the relative position of cities. The second one (Section 4.5) involves the location of firms but often neglects land. One of the fundamental challenges for future research is to merge these two structures to study the location of cities using land. In this perspective, one may wonder what the efficiency result of Proposition 4.3 becomes when the choice of locations by developers becomes critical.

Second, Section 4.2 shows how detailed economic analyses are able to uncover some of the forces leading to workers' agglomeration around a CBD. In particular, as argued in Chapter 1, appealing explicitly to imperfect competition has allowed a sharper grasp of the mechanisms at work. In such disaggregate models, instead of using ad hoc expressions, it is possible to evaluate the impact of a particular policy on all the agents and to design better corrective public policies. We believe that this kind of approach may also lead to better empirical models because they suggest what could be the relevant variables to consider. Although it is likely that black box models will remain useful in urban economics, not all black boxes are alike. We paraphrase George Orwell in *Animal Farm* by observing that all black boxes are black but some boxes are more black than others.

NOTES

1. This was already known by Desrousseaux (1964).
2. Both approaches have much to share with club theory à la Buchanan.
3. Note that a fourth approach has recently been suggested by McLaren (2000), although his purpose is not to deal with cities. The industrial structure of a given area, such as a region, may take the form of either many specialized firms or of a small number of vertically integrated firms. The pattern of organization is considered as the outcome of globalization and history. This should not come as a surprise because we know from Coase (1937) that one of the main reasons for firms to exist is the need they face to minimize transaction costs. Inasmuch as trade costs correspond to a particular type of transaction cost, globalization should be accompanied with a move toward more "outsourcing" and less vertical integration. Because firms tend to subcontract a great deal of intermediate activities and to concentrate on their core competencies (see, e.g., Milgrom and Roberts 1992, chap. 4), the "local" supply of goods and services becomes more important to them, thus making this trend relevant for the city structure.
4. In the same vein, once it is recognized that bilateral search often arises on the labor market, cities lead to lower search costs and, hence, to higher human capital and productivity (see, e.g., Acemoglu 1996).
5. As observed by Duranton and Puga (2000), the Chamberlinian and Smithian approaches have much in common with the Marshallian externalities of type (ii) and (iii) discussed in Section 1.3. Because externalities of type (iv) correspond to local public goods, it is preferable to restrict the concept of Marshallian externality to the informal sharing of information across agents through face-to-face communications, as discussed in Chapter 6.
6. For notational simplicity, we assume that the final sector uses no labor. It is fairly straightforward to add labor as an additional input when the production function is of the Cobb–Douglas type (Abdel–Rahman and Fujita 1990).
7. Recall that the CES function was first introduced in the economic literature as a production function by Arrow et al. (1961).
8. See Hayek (1988, chap. 8) for a similar idea that tends to run against the dominant paradigm.
9. This approach is consistent with the idea of Marshallian externalities as modeled by Henderson (1987, 1988, chap. 2), but here the market interaction leading to increasing returns is considered explicitly instead of being assumed through an ad hoc specification of an externality affecting the production function. For the comparison of the two approaches as well as for the determination of the equilibrium and optimum city size in the present context, see Fujita (1989, chap. 8).
10. The model presented here is the "counterpart" of the standard spatial competition model discussed in detail later (Section 4.5).
11. By choosing a technology midway between the two adjacent competitors, each firm is able to gain more market power and to relax wage competition. The equidistant configuration of technologies is likely to be an equilibrium outcome of a game in which firms would choose their technologies before their wages (see, e.g., Economides 1989 and Kats 1995).
12. When workers choose in which city to work and reside, it is reasonable to assume that they have incomplete information about job requirements in any particular city

(Hesley and Strange 1990). There is no need to make this assumption here because we focus on what happens within a single city.

13. Dealing with the choice of technologies by firms leads to unnecessary technical difficulties.
14. For analytical simplicity, it is assumed that the entry of a new firm leads the incumbents to relocate equidistantly along the circle.
15. We assume that each firm is equally shared by all the individuals in the economy. When each city is small compared with the whole population, the action of a firm affects its workers only negligibly. Because profits are zero in equilibrium, the residents of each city have no income other than wages.
16. Note that there is no congestion in commuting. Adding congestion would make the commuting cost even more convex.
17. Observe that this optimization problem does not necessarily have a solution. For that, some fairly mild assumptions must be imposed on preferences and the production function (Fujita 1989, Proposition 5.8). Furthermore, there may be more than one optimal solution. Hereafter, however, we assume the unique existence of the optimum. See Mirrlees (1995) for a sufficient condition guaranteeing uniqueness.
18. The most popular version of the Henry George theorem is associated with local public goods, as discussed in Chapter 5.
19. This implies that no coalition of workers can form to guarantee themselves a higher utility level. The corresponding allocation therefore belongs to the core of the economy. See Scotchmer and Wooders (1987) for a general approach to club formation.
20. There is a formal analogy between the results presented in this section and those obtained in another strand of the economics literature: labor–management versus entrepreneurial management. Specifically, Meade (1972) and Drèze (1974; 1985) have shown that both institutional systems yield the same long-run equilibria when there is free mobility of factors and perfect competition. Labor–management eliminates both profits and wages as guides to business decisions and assumes that firms maximize their value added per worker. This setting is therefore similar to our community model when members' utilities are identified with workers' income.
21. Note that the Henry George rule is reminiscent of the Groves–Clarke mechanism used in public good theory and developed around the same time.
22. Mirrlees (1995) has proposed an alternative but related model in which firms produce under increasing returns and are monopsonistic competitors on the labor market.
23. As in Section 4.3, city corporations are assumed to be equally shared by all the consumers.
24. According to Henderson (1987, 1997a), there is a substantial amount of empirical evidence showing that small and medium-sized cities tend to be specialized.
25. Housing stands for all the goods (such as general retailing, schools, housing services) that cannot be traded between cities because of prohibitive transportation costs.
26. In fact, Hotelling's analysis had been anticipated in several respects by Launhardt ([1885] 1993, chap. 9), who questioned the validity of the price-taking assumption by observing that firms dispersed over space have some market power over the customers situated in their vicinity, thus allowing them to manipulate their prices to their own advantage.

27. There lies the root of Hotelling's brilliant idea to restore continuity in the firm demands by considering a nonatomic distribution of consumers. A similar idea has been used much later in a different context by Aumann (1964).
28. The assumption of equidistant firms is the spatial counterpart of that of equally weighted varieties in the Dixit–Stiglitz model.
29. The assumption of a perfectly inelastic demand could be removed without affecting our main results but at the expense of longer algebraic developments.
30. Beckmann (1972b) has shown that the equilibrium price is lower when firms compete in a two-dimensional space rather than in one-dimensional space. This is because each firm now has six direct competitors (located at the vertices of a hexagon) instead of two, thus making price competition fiercer.
31. This is equivalent to the SPLP studied in Section 2.5 when there is a continuum of locations, facilities have identical costs across locations, and demand is uniformly distributed.
32. It should be kept in mind that the entry of a new firm is supposed to lead to a new equidistant configuration of firms. This is so if the equidistant configuration is a location equilibrium and if firms can freely move to the new locations. If the latter assumption does not hold, the number of incumbents at the free-entry equilibrium might be lower than the optimal number of firms (Eaton and Wooders 1985).
33. This approach to firms' and consumers' locational choice is analogous to the Cournot–Walras model used in general equilibrium with imperfect competition (Bonanno 1990). In this model, firms select quantities, and prices are then established at the Walrasian equilibrium of the corresponding exchange economy. Hence, firms are able to determine the demand functions relating the quantities they supply to the equilibrium market prices. Using these inverse demands, firms choose their outputs at the Cournot equilibrium. In the spatial setting, the locations of firms correspond to outputs, and the residential equilibrium, which is influenced by the locations chosen by firms, corresponds to the competitive equilibrium in the Cournot–Walras model.
34. Note that this constant cannot be determined within the model because of our assumptions of a fixed lot size and of no vacant land. When there is an arbitrarily small amount of vacant land, this constant is given by $\bar{R} = \max_{r \in C} \min_{j=1,\ldots,m} \{p_j + t|r - r_j|\}$. In any case, the value of this constant does not affect the results of this section.
35. If firms are assumed to be aware and able to manipulate the utility level, then a Nash equilibrium may fail to exist (Scotchmer 1985).
36. It can also be shown that the optimal configuration of locations may be sustained as a Nash equilibrium in which firms simultaneously choose price and location once firms are allowed to capitalize the land rent which they create (Asami et al. 1993).
37. Note that modern approaches to club theory assume in line with urban economics that (1) there is a continuum of agents and (2) each agent belongs to a finite number of clubs – here one city (Ellickson et al. 1999).
38. That cities are often associated with the presence of particular spatial inheterogeneities, such as rivers or harbors, may prevent the replicability of cities. Because coordinating urban activities is an especially hard task, replicability may also fail due to a deficit in city developers (Papageorgiou and Pines 2000).

5

Cities and the Public Sector

5.1 INTRODUCTION

When seeking a reason for the existence of cities, the one that comes most naturally to mind is the supply of public services. In large measure, this is dictated by historical considerations. For example, in medieval Europe cities were signified in two ways: a physical boundary (*the walled city*) and a legal status (*the democratized city*). Clearly, walling a city exhibits increasing returns and corresponds to a local public good whose supply is governed by size effects: the length of a circular wall is $2\pi r$, whereas the size of the corresponding area is πr^2; the ratio of the circumference to the area falls as the radius r increases, and thus a larger number of individuals may be defended at a lower average cost. In addition to its defensive purpose, the wall was also the symbol of the city's political autonomy, and the corporate freedom of the towns brought emancipation to individuals.[1] Historians agree that the specific legal status, which may itself be interpreted as a local public good, was a major criterion for identifying the city at least until the end of the Middle Ages (Bairoch 1985, chap. 1), if not later on. This clear-cut separation no longer exists. The legal status has been homogenized, except for minor exceptions, in most nations. Urban activities have gradually extended beyond the physical boundaries of the city to create suburbs, which are now very much part of the city considered as an economic agglomeration. As a result, the modern city is more dispersed and has fuzzy boundaries. However, the availability of local public goods remains a major ingredient of modern cities because the congregation of a large number of people facilitates the mutual provision of collective services that could not be obtained in isolation.[2]

A *pure public good* is collectively consumed by all members of a community such as a city or a nation (Samuelson 1954b). Its consumption is *nonrivalrous* in the sense that each individual's consumption does not subtract from any other individual's consumption of that good. A pure public good's benefits are also *nonexcludable* because, once the good is provided, it is virtually impossible to exclude any member of the community from the benefits. Hence, unlike private goods, public goods can serve unlimited numbers of consumers without having quantity or quality degraded through congestion or increased costs. In contrast to this somewhat extreme model, most public services suffer from congestion (e.g., too many people attending an exhibit). Furthermore, consuming the public good often involves traveling. If a public good is located in space, there is competition for the limited land close to the public good. Hence, the social cost increases with the number of users because higher transportation costs are required to use the public good. The literature following Tiebout (1956) and Buchanan (1965) has argued that both these effects compromise the "purity" of public goods and make them more similar to private goods. In this chapter, we consider this broader class of "impure" public goods.

It is well known that consumers have incentives not to reveal their true preferences regarding pure public goods because they cannot be excluded from their consumption. In this respect, it was Tiebout's merit to observe that many public services, such as police and fire protection, schools, hospitals, and stadiums, are "local." By migrating to the jurisdictions that respect their tastes in terms of goods and tax schemes, consumers reveal their preferences. According to Tiebout (1956, 420),

> Moving or failing to move replaces the usual market test of willingness to buy a good and reveals the consumer–voter's demand for public goods. Thus each locality has a revenue and an expenditure pattern that reflects the desires of its residents.

Thus, if each locality competes for consumers by providing its own package of public goods and taxes, competition among communities and "voting with feet" by consumers may lead to the efficient provision of local public goods.

For this to work, consumer mobility is necessary. However, there is another aspect that has been neglected in the standard literature on local public goods. That is, the choice of a particular community implies the choice of residence, which, in turn, involves land consumption. This fact has an important consequence for local public goods models, namely, *land capitalization*. Capitalization means that the price of land embodies the benefits and costs of public services incurred by the residents. Hence, capitalization provides a natural measure of social surplus or willingness to pay for an increase in local public goods. In fact, capitalization and consumer mobility are inextricably linked: because consumers can move from unattractive locations to attractive locations, land prices will adjust to compensate for the differences in attractiveness. In other words, through capitalization and consumer mobility, populations are endogenous to local policies.

Although the traditional spatial competition model has "location without land" (see Section 4.5), new local public goods models have "land without location" in the sense that there are no transport costs within the community. In

Tiebout's local public goods model, consumers are mobile in the sense that they choose what jurisdiction or location to occupy, but once there, their accessibility to public goods is irrelevant. Consumers cannot use the public goods of a neighboring locality even if those public goods are closer. The tying of consumer benefits to residency in the jurisdiction is essential to the success of the model; without it, the land value would not capture all the benefits of its policies. This assumption is most vividly met in the "islands" model of Stiglitz (1977) in which it is infeasible for a consumer on one island to consume the public goods of another.

In the context of local public goods, it is generally not desirable to increase the size of the city population indefinitely even though the per capita cost of the public good is decreasing with the number of users. Indeed, even when the public good is pure, the marginal social cost of a consumer, which is identical to the additional commuting cost, increases. Therefore, in the same spirit as in Chapter 4, *there is a trade-off between transport costs and the cost of supplying the public good.* In general, the city will have a finite optimal size, which is determined by maximizing the utility level of the residents.

In pointing out the analogy between private goods and local public goods, however, Tiebout did not specify the objective function of the jurisdictions. Much of the debate following his work has revolved around this question. According to several authors, including Arnott (1979), Kanemoto (1980, chap. 3), Berglas and Pines (1981), Hochman (1981), and Henderson (1977, chaps. 3 and 10; 1985), there is a missing agent in Tiebout's local public good setting, namely, a land developer who capitalizes the benefits of the public good in the land rent. In such an institutional context, competition between land developers may lead to the efficient provision of local public goods in a spatial economy. Indeed, jurisdictions, which are now identified with land developers, can profit by respecting their residents' tastes when the provision of public goods is capitalized into land prices. Thus, if capitalized land values are included in profits, jurisdictions have an incentive to organize their affairs efficiently.

In the real world, local public goods are often provided under the form of a public facility designed to provide a bundle of services to a community of consumers (Tiebout 1961; Teitz 1968). From the practical point of view, the importance of public infrastructure in shaping cities as well as the quality of life within them has been emphasized, with humor, by Teitz (1968, 36):

> Modern urban man is born in a publicly financed hospital, receives his education in a publicly supported school and university, spends a good part of his time travelling on publicly built transportation facilities, communicating through the post office or the quasi-public telephone system, drinks its public water, disposes of his garbage through the public removal system, reads his public library books, picnics in his public parks, is protected by its public police, fire, and health systems; eventually he dies, again in a hospital, and may even be buried in a public cemetery.

In an otherwise homogeneous space, the location of a public facility becomes the center of the city. This is because consumers must travel to this facility to enjoy the public services made available there.[3]

In Section 5.2, we study the optimal provision of local public goods in a system of cities populated with identical consumers. We show that, when the population size is such that the residents' common utility level is maximized, the cost of the public good is equal to the aggregate differential land rent within each city. Furthermore, if land developers take the common utility level in the economy as given, the first best optimum can be sustained as a free-entry equilibrium among land developers. In each city, the cost of the local public good is financed by the aggregate differential land rent. These results suggest that the rules governing the supply of local public goods are similar to those applicable to private goods, as discussed in Chapter 4. In other words, we have "Tiebout without politics" (Pines), or, as Henderson (1977, 72) put it:

> The existence of land developers seeking to maximize profits ensures that scale economy benefits of increasing city size versus commuting cost increases are traded off implicitly or explicitly to achieve optimal city size.

In the preceding approach it is assumed that there is enough land for each city to be developed as an "isolated city state" (Papageorgiou). At this point, it is therefore natural to ask *where* a local public good should be supplied when consumers are dispersed over the entire territory. Furthermore, we are also interested in checking what happens when the provision of public services is decided through a political process such as voting. This is a topic that has recently attracted attention in political economics. In Section 5.3, we consider a voting procedure in which consumers vote first for the number of facilities and then for their locations. As expected, each facility is set up at the middle point of its service area. Less expected is the result that, when the construction of these facilities is financed through a proportional income tax, voters tend to choose a number of facilities exceeding the efficient one. This suggests that the recourse to voting for choosing the system of public facilities fosters a proliferation of public infrastructure.

However, using more sophisticated taxation schemes enables one to sustain the optimum as a voting equilibrium. Specifically, we will show that the optimal tax scheme has the same profile as the land rent. More precisely, voting yields the optimum when the benefits associated with the proximity of facilities is capitalized in the land rent and when consumers are aware that they have to pay the corresponding rent. This shows, once more, that a perfectly competitive land market is a powerful device to achieve the first best optimum.

5.2 THE CITY AS A PUBLIC GOOD

Into the urban land use model considered in Section 3.3, we now introduce a third commodity in the consumers' utility function, that is, a *local public good*.

This good is made available to consumers through a facility located in the city. For simplicity, each city is to be formed within a one-dimensional space with unit land density. Consumers must bear some travel costs $T(r)$ to have access to the public service. In their attempt to reduce the access costs, consumers agglomerate around the place where the public facility is built in the same way as they do around the business district.[4] Let $g = g(G, N)$ be the quantity of public good, where G stands for public expenditure and N for the mass of users. If the local public good is *pure*, then g is independent of N and, without loss of generality, we may assume that $g = G$. If the local public good is *congestible*, then an additional consumer has a negative impact on the welfare of others, in which case g is a strictly decreasing function of the mass N of users.

The whole population in the economy is formed by $\mathcal{N}$ identical consumers whose income is Y. This income is earned in a perfectly competitive industry operating under constant returns to scale. The utility function of a consumer is

$$U[s, z, g(G, N)].$$

When the consumer resides at distance r from the city center, his budget constraint is given by

$$z + sR(r) = Y - T(r) - \theta(r)$$

in which $R(r)$ denotes the land rent prevailing at distance r and $\theta(r)$ any tax paid (or subsidy received) by a consumer at distance r. This tax depends only upon the consumer's location because consumers are identical up to their distance to the facility.

Suppose that the local public good is pure. Although the results we present below can be generalized to the case in which the lot size is variable, we find it convenient to assume that the lot size used by each consumer is fixed and normalized to one. Therefore, if N consumers live in the city, at the corresponding residential equilibrium they are evenly distributed around the city center over the interval $[-N/2, N/2]$. Let G be the level of public expenditure. Then, from the equality of the utility level across consumers and from the consumer budget constraint in which we set $s = 1$, a common level of composite good consumption, equilibrium consumption of the composite good z^*, and an equilibrium land rent $R^*(r)$ exist such that

$$z^* = Y - R^*(r) - T(r) - \theta(r) \quad r \in [-N/2, N/2]. \tag{5.1}$$

Thus, given G and $R^*(r)$, maximizing the utility of a consumer amounts to maximizing the consumption z^* as given by (5.1). In this case, evaluating the land rent at the urban fringe, we obtain

$$R^*(r) = T(N/2) + \theta(N/2) - T(r) - \theta(r) + R_A \quad r \in [-N/2, N/2], \tag{5.2}$$

where R_A is the agricultural land rent (or land opportunity cost).

In what follows, we first analyze the case of a single city and identify conditions under which confiscating the aggregate differential land rent is sufficient to finance the public good. We then consider the centralized provision of a local public good in a system of cities and discuss how the optimum can be decentralized by appealing to competition among land developers. Throughout this section, the amount of land available in the whole economy is assumed to be sufficiently large for the numbers of cities determined to be feasible.

5.2.1 The Henry George Theorem

Consider a group of individuals who choose to form an urban community to benefit from a local public good. In this section, the quantity of public good G is assumed to be fixed. Because individuals are identical, they delegate to a city government the task of maximizing their common utility level. To this end, the government first buys the land for the city from farmers at the agricultural rent R_A. Since the government knows that the competitive residential equilibrium is efficient (see Proposition 3.5), it may allow for a competitive land market to determine the consumer residential allocation and the consumption of the composite good within the city. Nevertheless, the government must find resources that allow it to finance the public good. To achieve its goal, the city government can confiscate the differential land rent created by the establishment of the public facility. In addition, the government may levy a tax $\theta(r) \geq 0$ that may vary with consumers' locations. The government is also entitled to choose the city population size N.

The city government understands that it is wasteful to have vacant land within the city and that the consumers must be symmetrically distributed about the public facility. Focusing on the right-hand side of the city, this implies that

$$R^*(r) - R_A \geq 0 \quad r \in [0, N/2] \tag{5.3}$$

and

$$R^*(N/2) = R_A. \tag{5.4}$$

Because each consumer's budget constraint is given by $Y = z^* + R^*(r) + T(r) + \theta(r)$, the equilibrium land rent must be such that

$$R^*(r) = Y - z^* - T(r) - \theta(r) \quad r \in [0, N/2]. \tag{5.5}$$

The city government has to solve the following problem:

$$\max_{N,\theta(.)} z^*$$

subject to the city budget constraint

$$2\int_0^{N/2} \theta(r)dr + 2\int_0^{N/2} [R^*(r) - R_A]dr \geq G \tag{5.6}$$

as well as to (5.3), (5.4), and (5.5).

Let ADR be the aggregate differential land rent when there are N consumers

$$ADR \equiv 2\int_0^{N/2} [R^*(r) - R_A]dr \tag{5.7}$$

and denote by $TTC(N)$ the total transportation costs incurred by the N consumers

$$TTC(N) = 2\int_0^{N/2} T(r)dr.$$

Substituting (5.5) into (5.6) and using the equality (no waste) yields

$$NY = Nz^* + TTC(N) + G + NR_A,$$

and thus

$$z^* = Y - \frac{TTC(N) + G + NR_A}{N}. \tag{5.8}$$

Accordingly, the optimal utility level corresponding to G is reached when the per capita cost $G/N + TTC(N)/N + R_A$ is minimized with respect to N. The trade-off discussed in the introduction should now be clear: if the population size rises, the per capita cost of the public good G/N decreases, but the per capita transportation cost $TTC(N)/N$ increases because the cost TTC is strictly increasing and strictly convex in N by Proposition 4.1.

Because (5.8) does not involve $\theta(\cdot)$, without loss of generality $\theta(r)$ may be set equal to zero for all r as long as (5.3) and (5.4) are met. Indeed, any positive or negative transfer is automatically reflected in the equilibrium land rent defined by (5.5). In this case, using (5.4), (5.5) becomes

$$R^*(r) = T(N/2) - T(r) + R_A$$

and, hence, ADR depends only upon N:

$$ADR(N) = 2\int_0^{N/2} [T(N/2) - T(r)]dr,$$

which is strictly increasing in N.

Furthermore, evaluating (5.5) at the urban fringe shows that

$$z^*(N) = Y - T(N/2) - R_A,$$

which strictly decreases with N. Consequently, maximizing $z^*(N)$ under the budget constraint (5.6), which is now rewritten $ADR(N) \geq G$, implies that the optimal population size $N^o(G)$ must satisfy the condition

$$ADR[N^o(G)] = G.$$

Thus, we have shown the following:

Proposition 5.1 *Given any level of expenditure on a pure public good in a city, the aggregate differential land rent equals public expenditure if the population size is chosen to maximize the utility level of the city's residents.*

In urban public finance, this result is known as the Henry George theorem based on the proposal for a confiscatory tax on pure rents made in 1879 by the American economist–crusader Henry George in his book *Progress and Poverty*.[5] It is worth stressing that Proposition 5.1 does not depend on the structure of preferences and holds regardless of the quantity of public good supplied within a city. This is to be contrasted to the standard equilibrium conditions for the efficient supply of a pure public good identified by Samuelson (1954), which requires knowing the marginal utility of the supplied public good across all consumers. Further, because $\theta(r) = 0$ for all r, *a single tax on land rent is sufficient to finance the public expenditure.* The land tax proposed by George also has the advantage of being levied on land, which is supplied inelastically so that no distortion is introduced in the price system. In practice, however, implementing such a taxation policy might be a difficult task.[6] Nevertheless, the idea is provocative, and we have already seen in Section 4.3.3 that similar mechanisms can be applied to firms providing private goods.

Furthermore, it follows from Proposition 5.1 and because $ADR(N)$ is a strictly increasing function of N that

$$ADR(N) \lesseqgtr G \quad \text{if and only if} \quad N \lesseqgtr N^o(G).$$

Stated differently, the aggregate differential land rent exceeds expenditure on the public good in a city with a population above the optimal size: too large a number of consumers leads to increasing land rent at each urban location. By contrast, expenditure on the public good exceeds the aggregate land rent in a city with a population below the optimal size: too small a number of consumers makes the land rent too low at each urban location. In this case, a tax is needed to finance the public good.

It can be shown that the Henry George theorem remains valid when the lot size is variable (hence, the population density now decreases as one moves away from the public facility) as well as in the presence of locational amenities (see Arnott and Stiglitz 1979 and Fujita 1989, chap. 6, for more details).

5.2.2 The Centralized Provision of a Local Public Good

Consider an economy-wide planner whose objective is to maximize the common individual utility level among the whole population $\mathcal{N}$ in the economy. The role of the planner is to determine the number of cities, their population size and corresponding residential area, the supply of public good in each city and the corresponding tax scheme, the allocation of consumers within each city, and their consumption of the composite good. Because all cities are identical, it is sufficient to focus on the representative city. As seen in the foregoing section, the planner allows for a competitive land market to determine the allocation of residential land to consumers and the consumption of the composite good within each city, whereas confiscating the aggregate differential land rent permits the financing of the public good.[7] In order to do so, he acquires the land needed for each city at the prevailing agricultural rent R_A. If N consumers choose to reside within the city, the planner knows that land will be used only for residential purpose and that consumers will distribute themselves over the segment $[-N/2, N/2]$. As a result, the planner has only to choose the population size N per city (or, equivalently, the number of cities $\mathcal{N}/N$) as well as the public expenditure G and the taxation scheme $\theta(\cdot)$ in each city. In the case of a pure local public good, the planner problem may then be written as follows:

$$\max_{G,N,\theta(.)} u = U(1, z, G)$$

subject to (5.3), (5.4), (5.5), and (5.6). Except for the choice of G, this problem is equivalent to the one considered in the preceding section.

Let $Z(u, G)$ be the unique solution to the equation $U(1, z, G) = u$. Then, using consumers' budget constraints, the city budget constraint (5.6) becomes

$$2\int_0^{N/2} [Y - Z(u, G) - T(r) - R_A]dr \geq G$$

in which the tax $\theta(\cdot)$ cancels out. The planner's problem may then be rewritten as follows:

$$\max_{G,N,\theta(.)} u = U(1, z, G)$$

subject to (5.3), (5.4), (5.5), and

$$2\int_0^{N/2} [Y - Z(u, G) - T(r) - R_A]dr \geq G. \tag{5.9}$$

As before, the value of $\theta(r)$ has no impact on the solution as long as the residential area is given by $[0, N/2]$, and thus we may choose $\theta(r) = 0$ for all r without loss of generality. Hence, the planner's problem reduces to

$$\max_{N,G} u = U(1, z, G)$$

subject to (5.3), (5.4), (5.5), and (5.9). Let (u^o, N^o, G^o) be the optimal solution.[8]

To investigate what the solution to this problem is, we first note that, when the reservation utility is u^o and when $G = G^o$, if the migration to the city were free, the equilibrium population of the representative city would be

$$N^*(u^o, G^o) = 2\mu\{r \geq 0; Y - Z(u^o, G^o) - T(r) - R_A \geq 0\},$$

where $\mu(\cdot)$ is the physical size of the corresponding residential area.[9] Set

$$\varepsilon \equiv N^*(u^o, G^o)/2 - N^o/2.$$

Because $T(r)$ is strictly increasing in distance and $Y - Z(u^o, G^o) - T(r) - R_A = 0$ when it is evaluated at $r = N^*(u^o, G^o)/2$, it can readily be verified that $\varepsilon \neq 0$ and (5.9) imply that

$$2\int_0^{N^*(u^o,G^o)/2} [Y - Z(u^o, G^o) - T(r) - R_A]dr$$
$$> 2\int_0^{N^o/2} [Y - Z(u^o, G^o) - T(r) - R_A]dr \geq G^o.$$

Therefore, since $Z(u, G^o)$ is continuous in u, these would exist $u' > u^o$ and $\delta > 0$ such that

$$2\int_0^{N^*(u',G^o)/2-\delta} [Y - Z(u', G^o) - T(r) - R_A]dr \geq G^o$$

and

$$Y - Z(u', G^o) - T(r) \geq R_A \qquad r \in [0, N^*(u', G^o)/2 - \delta]$$

with the equality holding at the urban fringe $N^*(u', G^o)/2 - \delta$. This would contradict the optimality of (u^o, N^o, G^o), and thus $\varepsilon = 0$. This means that the optimal population size N^o equals the equilibrium population size of the open city model in which $G = G^o$, and the reservation utility level is u^o.[10] Stated differently, we have

$$N^o = N^*(u^o, G^o). \tag{5.10}$$

Because ε must be 0 in the optimal solution, we have

$$2\int_0^{N^o/2} [Y - Z(u^o, G^o) - T(r) - R_A]dr = ADR^*(u^o, G^o).$$

Consequently, since (5.9) holds with equality at the optimum, it follows that

$$ADR^*(u^o, G^o) = G^o. \tag{5.11}$$

More generally, when the reservation utility is u and G is a public good supplied in the city, the aggregate differential land rent corresponding to the residential equilibrium of the open city model is such that

$$ADR^*(u, G) = 2\int_0^{N^*(u,G)/2} [Y - Z(u, G) - T(r) - R_A]dr, \tag{5.12}$$

where $N^*(u, G)$ denotes the equilibrium population size associated with u and G. Assume now G' exists such that

$$ADR^*(u^o, G') > G'.$$

It then follows from (5.12) that

$$ADR^*(u^o, G') \equiv 2\int_0^{N^*(u^o,G')/2} [Y - Z(u^o, G') - T(r) - R_{\text{A}}]dr > G'.$$

Again, $u' > u^*$ would exist such that

$$ADR^*(u', G') > G',$$

thus contradicting the optimality of (u^o, N^o, G^o). Therefore, it follows that

$$ADR^*(u^o, G) \le G \quad \text{for all } G \ge 0. \tag{5.13}$$

To sum up, if (u^o, N^o, G^o) is an optimal solution, the three conditions (5.10), (5.11), and (5.13) must be satisfied. Conversely, it can readily be verified that these three conditions are also sufficient for (u^o, N^o, G^o) to be an optimal solution to the planner's problem. In addition, (5.11) and (5.13) imply that

$$\frac{\partial ADR^*(u^o, G^o)}{\partial G} = 1. \tag{5.14}$$

Furthermore, using (5.12) and the envelope theorem, we obtain

$$\begin{aligned}\frac{\partial ADR^*(u^o, G^o)}{\partial G} &= -2\int_0^{N^*(u^o,G^o)/2} \left[\frac{\partial Z(u^o, G^o)}{\partial G}\right] dr, \\ &= -N^o \frac{\partial Z(u^o, G^o)}{\partial G}\end{aligned}$$

and thus

$$\frac{\partial ADR^*(u^o, G^o)}{\partial G} = -N^o \frac{\partial Z(u^o, G^o)}{\partial G} = 1. \tag{5.15}$$

This is equivalent to the standard Samuelsonian condition, which states that the optimal quantity of a public good is such that the sum of the marginal rates of substitution between the public good and the numéraire is equal to its marginal production cost (which is here equal to 1). In addition, (5.15) also implies that the marginal social value of the local public good is equal to the marginal increase in the aggregate differential land rent. This is a land capitalization rule (Starrett 1988, chap. 13).

Finally, provided that the optimal city size is sufficiently small relative to the whole population $\mathcal{N}$, the integer problem may be neglected, and the optimal number of cities is given by $\mathcal{N}/N^o$, where $N^o \equiv N^o(G^o)$.

5.2.3 The Supply of Local Public Goods by Land Developers

Following a well-established tradition in urban economics, we now assume a different institutional setting in which each city is developed by a profit-maximizing land developer. We now consider a market mechanism in which consumers are free to move into any city (N is endogenous to the city developer) and to choose the location they find appealing to them in the corresponding city (the residential equilibrium is also endogenous). Furthermore, consumers being identical, they must achieve the same utility level regardless of the city in which they live. Because many cities exist, each developer treats the utility level prevailing in the economy as given. Clearly, the assumption according to which the utility level is exogenous to each developer has a "competitive" flavor. It is reasonable, provided that the number of developers (cities) is large enough, for each consumer to consider his impact on consumers' welfare as negligible. Note that this assumption does not necessarily mean that a developer is able to observe the prevailing utility level. It just means that the developer believes (for whatever reason) that he cannot manipulate the reservation utility.

In such a context, the developer's policy is, therefore, to attract some consumers by supplying them with a local public good, taking their utility level as given. When consumers decide to reside in a city, its developer may charge them a fee (which may be positive or negative) $\theta(r)$ that may vary with the distance to the facility (here the city center). As in Chapter 4, it is assumed that the developer is able to anticipate the residential equilibrium corresponding to his policy on G and $\theta(.)$, taking the utility level u as given.

The developer's profit is equal to the aggregate differential land rent plus the total fee collected from the residents minus the expenditure on the local public good. Hence, the maximization problem of a developer can be written as follows:

$$\max_{G,\theta(.)} \Pi[G,\theta(\cdot);u] = 2\int_{X_R^*}[R^*(r) - R_{\mathrm{A}}]dr + 2\int_{X_R^*}\theta(r)dr - G, \tag{5.16}$$

where $X_R^* = \{r \geq 0; R^*(r) \geq R_{\mathrm{A}}\}$ denotes the equilibrium residential area, whereas the corresponding equilibrium land rent $R^*(r)$ is given by

$$R^*(r) = Y - Z(u,G) - T(r) - \theta(r). \tag{5.17}$$

Substituting (5.17) into (5.16), we obtain

$$\begin{aligned}\Pi[G,\theta(\cdot);u] &= 2\int_{X_R^*}[Y - Z(u,G) - T(r) - \theta(r) - R_{\mathrm{A}}]dr \\ &\quad + 2\int_{X_R^*}\theta(r)dr - G \\ &= 2\int_{X_R^*}[Y - Z(u,G) - T(r) - R_{\mathrm{A}}]dr - G.\end{aligned} \tag{5.18}$$

For a given value of G, it follows from (5.18) that choosing $\theta(r)$ is equivalent to choosing the residential area X_R^*. Clearly, the profit-maximizing residential area X_R^* is the domain over which the willingness to pay for land exceeds its opportunity cost, that is,

$$X_R^* = \{r \geq 0; Y - Z(u, G) - T(r) \geq R_{\mathrm{A}}\}. \tag{5.19}$$

Hence, as before, the profit-maximizing value of $\theta(r)$ can be set equal to zero.

Because consumers maximize utility, they move to the developer's city up to the point in which the utility prevailing there is equal to u. Let $N^*(u, G)$ be the equilibrium population corresponding to G (as defined in Section 5.2.2) so that $X_R^* = [0, N^*(u, G)/2]$.Using (5.12), we may rewrite a developers' profit as follows:

$$\begin{aligned}\Pi(G; u) &= 2\int_0^{N^*(u,G)/2} [Y - Z(u, G) - T(r) - R_{\mathrm{A}}]dr - G \\ &= ADR[u, N^*(u, G)] - G.\end{aligned} \tag{5.20}$$

Consequently, each developer maximizes $\Pi(G; u)$ with respect to G when the utility level is u.

If there is free entry and exit (see also Section 4.3.3), then developers enter (leave) the city market as long as potential profits obtained by developing a city are positive (negative). During this process, the utility level u, the quantity of local public good G, and the population N of each city vary. When the long-run equilibrium is reached, profits (5.20) are zero:

$$ADR[u^*, N^*(u^*, G^*)] = G^*, \tag{5.21}$$

and thus profit-maximizing behavior also implies that

$$ADR[u^*, N^*(u^*, G^*)] \leq G \quad \text{for all } G \geq 0. \tag{5.22}$$

Finally, the equilibrium population N^* satisfies

$$N^* = N^*(u^*, G^*). \tag{5.23}$$

Conditions (5.21), (5.22), and (5.23) are necessary and sufficient for the market outcome (u^*, N^*, G^*). They are identical to (5.10), (5.11), and (5.13) determining the optimum. Hence, it follows that $u^* = u^o$, $G^* = G^o$, and $N^* = N^o$. Consequently, the number of cities arising at the market equilibrium $\mathcal{N}/N^*$ is also equal to the optimal number of cities $\mathcal{N}/N^o$.

Because the Henry George theorem holds in each city, the aggregate differential land rent collected by each developer allows him to finance exactly the efficient provision of local public good, which is a result comparable to that obtained in Section 4.4.3 for private goods.

In addition, we have the condition

$$\frac{\partial ADR^*(u^*, G^*)}{\partial G} = 1,$$

which is similar to (5.14).

All this analysis may be summarized as follows:

Proposition 5.2 *When the local public good is pure, an urban system is efficient if and only if it is a free-entry equilibrium of the city market. At both outcomes, the public good in each city is solely financed by the aggregate differential land rent.*

Thus, if the entrepreneur is brought back under the form of a land developer, the supply of a local public good seems to obey rules similar to those governing the production of a private good. This process is reminiscent of that described in Section 4.3.3, although the reasons for the emergence of urban agglomerations differ.

5.2.4 The Case of a Congestible Public Good

Consider now a public good such that an additional resident has a negative impact on the consumption of this good by the incumbents. This is probably more realistic than the case studied in the preceding section because most public facilities have a maximum capacity. According to Buchanan (1965), Berglas (1976), and Scotchmer and Wooders (1987), congestion would be sufficient to foster the decentralized provision of public services by clubs, internalizing the trade-off between financing and congestion. For this to occur, each club must be able to charge a fee equal to the congestion cost generated by an additional user and imposed to all users. Charging such a fee then allows us to finance a congestible public good and provides the right incentives to choose the optimal quantity for the users of the good. In this context, a city can be viewed as a "consumption club" in much the same way as a city is considered a "production club" in Section 4.3.

In each city, the land developer (or the local jurisdiction) now maximizes the aggregate differential land rent plus the total fee charged to all users of the congestible public good made available by the developer minus the cost of this good (see Berglas and Pines 1981 and Fujita 1989, chap. 6, for more details). In particular, it can be shown that $\theta(r)$ is no longer equal to zero but to an admission fee $\theta^o(N^o) > 0$ for all r. Both types of tax must be combined to finance the public good, implying that the Henry George theorem is amended in the following way: When the city size is optimal (N^o), public expenditure equals the aggregate differential land rent plus the optimal user fee $\theta(N^o)$ collected from all the users:

$$G = ADR(N^o) + N^o\theta(N^o).$$

Hence, once we account for the spatial setting, a Pigovian tax falls short of the provision cost, but the deficit is just equal to the differential land rent. Hence, the decentralization of a local, but congested, public good is still possible. Observe, finally, that the optimal fee depends here only upon the mass of consumers patronizing a facility and not upon their residential locations.

5.2.5 Potential Limits of Land Capitalization

Appealing as the capitalization argument seems, it needs qualification.

1. For the result above to hold, the city boundaries must be variable. Indeed, Proposition 5.2 no longer holds when borders are determined by administrative rules. Revising urban and regional borders certainly goes against the customary habit of regarding administrative boundaries as permanent. This rigidity responds, at least partly, to the individuals' need to belong to a lasting community whose geographic contours must remain stable. Agents involved, therefore, need to handle a new socio-economic trade-off. In any event, Henderson (1985) observed that such a process of revision is not unusual. In the United States, the growth of many cities took place through the annexation (and detachment) of smaller spatial entities. A similar process took place in Europe during the two main waves of urbanization in the twelfth and nineteenth centuries (Pirenne 1925; Bairoch 1985, chap. 10).[11]

   Furthermore, for land-value maximization to yield efficiency, the city must include all the beneficiaries of its fiscal policy within its border. Hence, to avoid uncounted *spillovers*, cities must be sufficiently large – something that may require the annexation of suburban communities. This might not be easy to do, for local governments and communities will resist precisely because their autonomy allows them to free ride on the city's provision of public goods.

   Finally, we also encounter neighborhood public goods within a city (Fujita 1989, chap. 6). These goods are to the city what local public goods are to the nation in the sense that they have a utility only to the residents who live in a particular urban neighborhood of the city. The land rent prevailing in a neighborhood increases with the provision of such goods, and the principles described above remain valid provided that the supplier can recover the increment in the land rent profile created by his decision to offer a particular neighborhood public good. However, implementing such a policy at a low spatial level seems to be fairly problematic because the service areas tend to be fuzzy and spillovers are likely to be important.
2. Each city must be small relative to the total population. In this case, the prices of land in other cities, hence utility, will be almost unresponsive to one city's change in fiscal policy. If the city is large relative to the economy, the competitive hypothesis may fail to hold. Indeed,

as pointed out in the introduction, capitalization and consumer mobility go hand in hand. Land prices rise in a city because more public goods attract residents from elsewhere, which pushes up demand for land and increases the population of the city. Meanwhile, as residents leave the other cities, the land prices there fall, providing the remaining residents more utility than they had before. In this way, utility is "exported" from the city that increased its public goods to the other cities. As a result, with a small number of cities, the utility-taking assumption is no longer tenable, thus making competition among land developers strategic.

Because of this utility effect, the increase in land value in the city that has increased its supply of public goods may underestimate willingness to pay for the public goods, and in equilibrium public goods may be underprovided. When the utility effect described in the preceding paragraph is strong enough, underprovision arises because firms compete by using only public goods. However, as Scotchmer (1986) has shown, a city can do potentially better if it has two instruments to govern these two effects. If cities can charge a head tax to control migration ($\theta(N) > 0$), public goods will be provided efficiently within each city, but the allocation of population among cities may not be optimal.

3. We have assumed so far that a single facility is able to supply all public services to the city's residents. Instead, one should expect different local (pure) public goods to be supplied by different facilities that are not necessarily located together. Indeed, the efficient number and locations of facilities generally differ across goods and services. As a result, each consumer patronizes different public facilities and, therefore, belongs to different service areas. Hence, the problem is now to find a way to distribute the differential land rent among the various local public goods consumed by the residents. Because each type of facility is likely to have a specific service area, the relative position of a consumer with respect to the nearest facility supplying each type of good varies with this consumer's location. In other words, the "contribution" of a particular facility to the differential land rent changes with the residential location.

This is not the end of the story. The problem of giving the right incentives to each supplier to make the efficient decisions regarding the location and size of his facility remains. In the aggregate, the Henry George theorem holds true. But the rent-sharing rule must relate the proceeds to the actions taken by each supplier of a local public good. More precisely, the supplier's revenue must exactly reflect the net social benefit of his decision for all his patrons operating within the various facilities. For example, if some fixed rule is applied, each supplier receives a certain fraction of the differential land rent, but then his marginal profit generally differs from the net social marginal benefit.

As shown by Hochman, Pines, and Thisse (1995), geography has a major implication for the socially desirable institutional structure of local governments: the efficient provision of local public goods must be decentralized through metropolitan governments that supply the whole range of local public goods over some appropriate territories. For this to hold, the area monitored by the metropolitan government must be sufficiently large. To avoid the degenerate situation in which the whole economy is used as a reference, one may assume that the metropolitan region should be small enough that all trips made within this territory to consume public services both originate and end there (Hochman et al., 1995). This permits the corresponding territory to include an integral number of each service area. In this way, the rent-sharing problem is solved because the metropolitan government internalizes all the benefits generated by the facilities under its control.

But practically, this implies that the corresponding units will often be large. One then runs the risk that the metropolitan governments will behave strategically. This would invalidate the extension of Proposition 5.2 to the multiservice case. Furthermore, the setting of large local governments could generate other well-known inefficiencies.

4. Finally, the assumption of identical consumers is obviously very strong. When there are several types of consumers, different types of cities, involving different mixtures of individual types, are likely to emerge (see, e.g., Scotchmer 1994, 2001). Hence, in this case, even when the economy is large, it is far from being obvious that land developers will be utility-takers. Indeed, it is likely that they will know that the change in one particular city has a direct and significant impact only upon a few other cities. In other words, competition among land developers becomes localized as in Section 4.5.

5.3 THE NUMBER AND SIZE OF CITIES UNDER POLITICS

The preceding section has shown that consumers want to be organized into communities whose number is governed by the existence of a trade-off between transportation costs and the cost of the public good. Furthermore, this trade-off is solved optimally when the provision of the public services is capitalized into the land rent and when developers are able to collect the differential land rent they create. However, in this setting, a city's area is not constrained by the other cities, and each one freely chooses how much land to use. One may wonder what our results become when consumers are continuously dispersed over the entire space within a given territory while their loctions are fixed.[12] As in Section 5.2, the number and size of cities are endogenous, but they occupy the whole territory whose size is fixed; hence, cities are connected through endogenous borders. It seems reasonable to think of this problem as one in which consumers are asked to express their preferences about local public goods through a political process such as voting. This subject has been addressed by Cremer, de Kerchove, and Thisse (1983; 1985) and, more recently, by Alesina and Spolaore (1997) in a slightly different context.

Because legal factors may prevent the use of other taxes or political difficulties may prevent the implementation of specific taxes,[13] the public good is assumed to be financed through an income tax. Thus, consumers voting on the number and location of public facilities are aware that they are both users of public goods and taxpayers. As users, they would like to have a public facility as close as possible to their residence, thus fostering the proliferation of such facilities. However, as taxpayers, they also understand that their tax bill will increase with the number of facilities, thus leading to the reduction in the supply of facilities. Hence, each voter internalizes the trade-off but does not take into account the impact of his decision upon the others.

5.3.1 The Political Economics of Community Formation

Voting is modeled as a two-stage process in which consumers first vote on the number of facilities and then on their locations.[14] This division into stages parallels the actions of consumers in choosing the number of facilities (the first stage) before making decisions regarding their location (the second stage). When choosing the number of facilities, voters anticipate the locations of the corresponding facilities. For simplicity, we make the following assumptions: (1) the local public good is pure and supplied in a fixed quantity so that its cost is given and equal to G for each facility and (2) transport costs are linear in distance, that is, $T(r) = tr$.

Space is described by a linear segment with length l. Consumers are evenly distributed over this segment, perhaps because they consume a fixed lot size, and the uniform density is equal to a constant n. If M facilities are built, the global budget constraint under a proportional income tax θ is given by[15]

$$(\theta Y)(nl) = MG. \tag{5.24}$$

Consider a locational configuration of M facilities $\mathbf{y}_M = (y_1, \ldots, y_M)$ such that facilities are placed at $0 \leq y_1 < \cdots < y_M \leq l$. If a consumer at $x \in [0, l]$ patronizes the facility located at y_i, his budget constraint is

$$Y = z + t|x - y_i| + \theta Y.$$

Note that consumers' locations are fixed so that there is no land market and hence no land rent. Because the quantity of public good is the same at each facility, each consumer patronizes the nearest facility. Since G is also fixed, a consumer located at x maximizes his utility $U(1, z, G)$ (see Section 5.2) if and only if the consumer at x maximizes consumption of the composite good, which is given by

$$z(x; M, \mathbf{y}_M) = Y(1-\theta) - \min_{i=1,\dots,M} t|x - y_i|$$
$$= Y - \min_{i=1,\dots,M} t|x - y_i| - \frac{MG}{nl} \tag{5.25}$$

in which we have used (5.24).

Let us now describe the voting procedure. In the second stage, the number of facilities is known to the voters as the outcome of the first-stage voting game. The utility of a consumer, therefore, depends on the location of the nearest facility, and the utility level decreases as the distance to this facility increases. This yields a voting subgame whose outcome is defined as a *Condorcet equilibrium* (i.e., a locational configuration such that no other locational configuration with the same number of facilities is strictly preferred by a strict majority of voters). When the number of facilities is M, the Condorcet equilibrium is denoted by $\mathbf{y}_M^* = (y_1^*, \dots, y_M^*)$.

To illustrate how this works, consider two configurations $\mathbf{y}_1$ and $\mathbf{y}_2$ with M facilities each. Then, if the mass of consumers who strictly prefer $\mathbf{y}_1$ to $\mathbf{y}_2$ exceeds the mass of consumers who strictly prefer $\mathbf{y}_2$ to $\mathbf{y}_1$, $\mathbf{y}_1$ is collectively chosen. Because the population density is uniform, the size of the area in which consumers strictly prefer $\mathbf{y}_1$ to $\mathbf{y}_2$ is larger than or equal to the size of area in which consumers strictly prefer $\mathbf{y}_2$ to $\mathbf{y}_1$:

$$\mu\{x; z(x; M, \mathbf{y}_1) > z(x; M, \mathbf{y}_2)\} \geq \mu\{x; z(x; M, \mathbf{y}_1) < z(x; M, \mathbf{y}_2)\},$$

where μ measures the size of the corresponding area.

We can now study the first-stage voting game in which consumers choose the number of facilities. In doing so, they anticipate the outcome of the voting subgame induced by their choice. Hence, the utility of a consumer at x is given by the utility achieved in the second stage, that is, by (5.25) in which $\mathbf{y}_M$ is replaced by the Condorcet equilibrium $\mathbf{y}_M^*$. This game is solved at the number of facilities such that no other (integer) number is strictly preferred by a strict majority of voters. This works as follows. If two numbers of facilities, M_1 and M_2, are proposed, M_1 is chosen if the mass of consumers who strictly prefer M_1 to M_2 is larger than or equal to the mass of consumers who strictly prefer M_2 to M_1. Formally,

$$\mu\left\{x; z\left(x; M_1, \mathbf{y}_{M_1}^*\right) > z\left(x; M_2, \mathbf{y}_{M_2}^*\right)\right\} \geq \mu\left\{x; z\left(x; M_1, \mathbf{y}_{M_1}^*\right) < z\left(x; M_2, \mathbf{y}_{M_2}^*\right)\right\}.$$

The Condorcet equilibrium of the first-stage voting game is denoted M^*.

The final outcome of the voting procedure, called a *subgame perfect Condorcet equilibrium*, is such that the following two conditions are met: (1) for each integer M, there are more consumers who strictly prefer the configuration $\mathbf{y}_M^*$ to any other configuration $\mathbf{y}_M$ with M facilities, and (2) for all $M \neq M^*$, there are more consumers who strictly prefer M^* to M than consumers who strictly prefer M^* to M. This is denoted by $(M^*, \mathbf{y}_M^*)$.

As usual, the game is solved by backward induction. In the second stage, the utility of a consumer at x reduces to

$$z(x; M, \mathbf{y}_M) = Y - \min_{i=1,\ldots,M} t|x - y_i| - \frac{MG}{nl},$$

which is single-peaked about x. This problem is, therefore, reminiscent of the median voter principle (see, e.g., Mueller 1979). However, this principle does not apply here because M items – the locations of facilities – instead of one are to be chosen by the voters. This implies that we need a specific result, which we state in Proposition 5.3 (the proof is contained in Part A of the chapter appendix).

Proposition 5.3 *Assume a proportional income tax. Then, for any given* M, *the equidistant configuration* $\mathbf{y}_M^* = (l/2M, \ldots, (2M-1)l/2M)$ *is the unique Condorcet equilibrium of the second-stage voting game.*

The intuition behind this result is as follows. Because each facility is at the middle of its service area, its location is the median of its consumer distribution. It is as if the median voter principle had been applied to each facility conditional on its respective service area. Let us show how the proof works when $M = 2$. If both facilities are located outside (inside) the first and third quartiles, then a majority of consumers are located on both sides of the center (in the two hinterlands generated by the facility's locations) who strictly prefer $(l/4, 3l/4)$ to the status quo. Assume now that facility 1 (2) is located between 0 and $l/4$ ($l/2$ and $3l/4$). Then, all consumers between $(y_1 + l/4)/2$ and $l/2$ strictly prefer a configuration with a facility at $l/4$. Similarly, all consumers between $(y_2 + 3l/4)/2$ and l strictly prefer a configuration with a facility at $3l/4$ (see Figure 5.1 for an illustration). Adding these two numbers, we obtain

$$\frac{n}{2}\left(l - y_1 - \frac{l}{4} + 2l - y_2 - \frac{3l}{4}\right) = \frac{n}{2}(2l - y_1 - y_2) > \frac{nl}{2}$$

because $y_1 < l/4$ and $y_2 < 3l/4$.

It should be clear that the equidistant configuration is also the one that minimizes total transport costs (see Part B of the chapter appendix for a proof). Hence, voting and planning yield the same outcome when the number of facilities is fixed, as in the case when individuals vote for a single item.

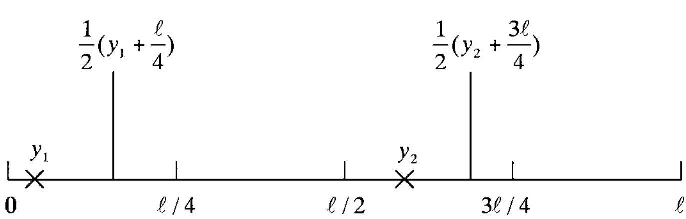


Figure 5.1: Comparing (y_1, y_2) to $(l/4, 3l/4)$.

Consider now the first stage. Consumers now vote on the number of facilities conditionally upon the locations obtained in the second stage. Hence, the utility of a consumer at x becomes

$$z(x; M, \mathbf{y}_M^*) = Y - \min_{i=1,\ldots,M} t|x - y_i^*| - \frac{MG}{nl},$$

where

$$y_i^* = (2i - 1)l/2M \qquad i = 1, \ldots, M.$$

This utility function is not single-peaked in M. Indeed, as M increases from 1 to some large integer, $z(x; M, \mathbf{y}_M^*)$ exhibits ups and downs corresponding to variations in the distance to the nearest facility. Consequently, a Condorcet equilibrium may not exist for some parameter configurations.

However, we have the following result whose proof is given in Part C of the chapter appendix.

Proposition 5.4 *Assume a proportional income tax. If*

$$G \geq \frac{(nl)(tl)}{40} \tag{5.26}$$

then there always exists a unique Condorcet equilibrium of the first-stage voting game given by the largest integer M satisfying the inequality

$$M(M-1) < \frac{(nl)(tl)}{2G}. \tag{5.27}$$

Accordingly, a subgame perfect Condorcet equilibrium exists when the cost G of a public facility is sufficiently large compared with the transport cost tl, the market size, nl, or both. Stated differently, condition (5.26) is likely to be satisfied when the territory or the population is small. It also follows from (5.26) and (5.27) that the equilibrium outcome M^* is such that it does not exceed 4.

One may wonder what happens when (5.26) does not hold. When $M = 5$, a simple calculation shows that $M = 3$ is strictly preferred to $M = 5$ by more than half of the voters. The same holds for any odd number M larger than 5, which can be shown to be defeated by $M - 2$. The reason is as follows. When M is odd, a facility is always established at the center of the segment, and consumers around it strictly prefer the equidistant configuration with $M - 2$ facilities to that with M facilities because the tax bill is lower. Consequently, there are only $M - 1$ groups of consumers located respectively around the remaining $M - 1$ facilities who strictly prefer M to $M - 2$. However, when $M \geq 5$, these groups are too small to form a majority in favor of M. Hence, in general, M is not a Condorcet equilibrium when M exceeds 4.

However, if we define a *local Condorcet equilibrium* for the first-stage voting game as a number defeating $M - 1$ and $M + 1$, we can obtain (see Part D of the chapter appendix for a proof) the following:

Proposition 5.5 *Assume a proportional income tax. Let $\bar{M}$ be the largest integer for which the inequality*

$$M(M-1) < \frac{(nl)(tl)}{2G} \tag{5.28}$$

holds; *then*, $\bar{M}$ *is the unique local Condorcet equilibrium of the first-stage voting game.*

The uniqueness of the local equilibrium endows the number $\bar{M}$ with some degree of stability. In addition, as shown by the proof of Proposition 5.5, the number $\bar{M}$ is also the single surviving outcome of the following sequential choice process. Consumers are first asked to vote between one or two facilities. Then, they are asked to choose between the winner and three facilities. The procedure is iterated until $\bar{M}$ is reached. Because $\bar{M}$ defeats any number $M > \bar{M}$, as shown in the proof of Proposition 5.5, it is indeed the only surviving outcome. Finally, when a Condorcet equilibrium of the first-stage voting game exists, it is also given by $\bar{M}$. All these results lead us to retain $(\bar{M}, \mathbf{y}^*_M)$ as the outcome of the voting procedure for all values of G.

Hence, given (5.28), we see that the voting number of cities increases as the cost of the public good (G) decreases, the transport cost (tl) increases, and the population size (nl) increases. Correspondingly, the size of each city decreases.

5.3.2 Is There an Oversupply of Public Facilities?

In the present setting, the efficient outcome is obtained by minimizing the social cost defined as the sum of fixed costs and transport costs. In view of the discussion in the preceding section, we see that, for a given value of M, the voting configuration is identical to the efficient one: they are both equidistant.

However, the numbers of facilities generally differ according to the decision-making procedure selected. Indeed, given (5.25), the efficient number of facilities minimizes the social cost defined as follows:

$$\begin{aligned} C(M) &= MG + 2M \int_0^{l/2M} ntr dr \\ &= MG + ntl^2/4M. \end{aligned} \tag{5.29}$$

Given the shape of $C(M)$, it can readily be verified that the efficient number of public facilities is the largest integer M^o that satisfies the condition (see also Section 4.5.2):

$$M(M-1) < \frac{(nl)(tl)}{4G}. \tag{5.30}$$

Comparing (5.28) and (5.30), we immediately obtain the following:

Proposition 5.6 *Assume a proportional income tax. Voting tends to set a number of facilities exceeding the number that a planner maximizing total welfare would choose.*

Hence, voting fosters an excessive number of cities as well as an oversupply in public infrastructure very much as the market does in the case of a private good (see Section 4.5.2). The discrepancy appears to be especially large when the public good cost is large and transport costs are low. In addition, each city is too small.

In particular, the divergence between $\bar{M}$ and M^o can be made more visible when these numbers are large in as much as $M(M-1)$ is almost equal to M^2. In this case, $\bar{M}$ is approximately equal to $\sqrt{2}M^o$, which means an increase of 40% in the public budget.

Somewhat surprisingly, the voting procedure yields the number of facilities that would be chosen by a Rawlsian planner who maximizes welfare of the worst-off consumer. Indeed, because the Rawlsian configuration is equidistant, the maximum distance covered by a consumer when there are M facilities equals $l/2M$. Hence, the social welfare function of such a planner is given by

$$Y - \frac{tl}{2M} - \frac{MG}{nl}. \tag{5.31}$$

Clearly, this function is maximized when M is the largest integer satisfying inequality (5.28). Recall that the outcome of voting is generally associated with the preferences of the median voter. In the community formation problem, however, the pivotal voter is the most extreme voter (when the public good is financed by an income tax). This can be explained as follows. The preferences of the individuals in the hinterlands $[0, l/2M]$ and $[(2M-1)l/2M]$ are determinants for the voting outcome. Furthermore, those individuals vote in the same manner. Accordingly, the voting number of facilities must be such that the consumers at 0 (or at l) strictly prefer $\bar{M}$ to $\bar{M}-1$ and $\bar{M}+1$. But these consumers belong to the set of worse-off ones in any equidistant configuration, and thus their best choice is socially desirable under Rawlsian optimality.

Hence, we have the following:

Proposition 5.7 *Assume a proportional income tax. The voting outcome coincides with the Rawlsian planning solution.*

This proposition may also be used to understand the intuition that stands behind Proposition 5.6. At the socially optimal number of facilities M^o, consumers' average utility is maximized, and this utility level is reached by the consumers located at a distance $l/4M^o$ from their nearest facility. As a result, the utility of a consumer located at distance $l/2M^o$ is much lower because he has to cover twice the average distance, which gives him an incentive to vote for a larger number of facilities. Since we just saw that these consumers are critical for the outcome of the voting process, it becomes simple to understand why voting yields a number of facilities larger than the efficient one.

Finally, a more general approach can follow once it is understood that (5.29) and (5.31) are the two polar cases of the social welfare function given by

$$\left\{ \int_0^l \left[\min_{i=1,\ldots,M} t|x - y_i| + \frac{MG}{nl} \right]^\alpha dx \right\}^{1/\alpha} \tag{5.32}$$

in which $\alpha \geq 1$ is a measure of the degree of aversion toward inequality (Papageorgiou 1977). Minimizing this function can be shown to yield the following results: (1) for any given value of M, facilities are equidistant, and (2) the socially optimal number of facilities is a nondecreasing function of α (Cremer et al. 1985).

As a consequence, the answer to the question raised in the title of this section varies with the nature of the planner's objective. If α is small, so that efficiency considerations are predominant, then voting fosters excess capacity and overtaxing. However, the discrepancy narrows as the population density (n) or the size of the territory (l) increases, whereas it enlarges when the cost of the public good (G) or the transport rate (t) decreases. In other words, when the economy is large in terms of either its population or its physical size, the voting outcome is not very different from the efficient one. By contrast, the gap between the two outcomes becomes significant when the transport rate is sufficiently low.

On the contrary, if α is large, so that equity considerations drive the choice made by the planner, then voting is socially desirable because it yields the socially optimal number of facilities. As a result, a trade-off exists between efficiency and spatial equity in the formation of public communities: spatial equity leads to a larger number of urban regions and a bigger public budget than efficiency.

When α increase from 1 to ∞, then the discrepancy between the voting outcome and the socially optimum one decreases.

5.3.3 The Role of Land Capitalization

Thus far, consumer locations have been assumed to be fixed so that there is no land market. In contrast, when consumers can move, they compete for land and pay a land rent. We want to reconsider the voting problem addressed in Section 5.3.1 in the presence of a land market. For this, we assume without loss of generality that $n = 1$ and, hence, the total number of consumers is l. We also assume that each consumer uses a unit lot size so that the equilibrium distribution of consumers is uniform.

The differential land rent arises here because of consumers' proximity to their nearest facility. In other words, if M facilities are placed at

$0 \leq y_1 < \cdots < y_M \leq l$, the bid rent of a consumer at $x \in [0, l]$ is given by

$$\Psi(x) = Y - z^* - \min_{i=1,\ldots,M} t|x - y_i| - \theta(x),$$

where z^* is the common equilibrium consumption of the composite good at the residential equilibrium corresponding to $(M, \mathbf{y}_M, \theta)$, and $\theta(x)$ is the tax paid (or the subsidy received) by this consumer. The equilibrium land rent is such that

$$R^*(x) = Y - z^* - \min_{i=1,\ldots,M} t|x - y_i| - \theta(x) \geq 0 \quad x \in [0, l]. \tag{5.33}$$

Furthermore, the total budget constraint is such that

$$\int_0^l R^*(x)dx + \int_0^l \theta(x)dx = MG.$$

Using (5.33), we can transform this constraint to

$$\int_0^l [Y - z^* - \min_{i=1,\ldots,M} t|x - y_i|]dx = MG \tag{5.34}$$

in which $\theta(x)$ cancels out. In turn, (5.34) yields

$$z^*(M, \mathbf{y}_M) \equiv z^* = Y - \frac{\int_0^l [\min_{i=1,\ldots,M} t|x - y_i|]dx + MG}{l}.$$

As a result, consumers cast their votes in order to maximize z^* subject to (5.34). Therefore, following the same voting procedure as in Section 5.3.1, we see that, in the second-stage voting subgame induced by M, all consumers agree to select the equidistant configuration because it minimizes total transport costs, thus maximizing consumers' equilibrium consumption of the composite good. Similarly, in the first stage, given the equilibrium locations resulting from the second stage, consumers unanimously choose the number of facilities minimizing total cost $C(M)$ given by (5.29).

Given (5.33), we have

$$R^*(x) + \theta(x) = Y - z^* - \min_{i=1,\ldots,M} t|x - y_i|.$$

Setting

$$R^*(x; M, \mathbf{y}_M) \equiv R^*(x) + \theta(x),$$

we obtain

$$R^*(x; M, \mathbf{y}_M) = Y - z^* - \min_{i=1,\ldots,M} t|x - y_i| \qquad x \in [0, l].$$

For the foregoing analysis, the respective share of the land rent, $R^*(x)$, and tax, $\theta(x)$, is immaterial. Consequently, we may focus on $R^*(x; M, \mathbf{y}_M)$, which we call the *land quasi-rent*.[16] Hence, in equilibrium, we have

$$\int_0^l R^*(x; M^*, \mathbf{y}^*_{M^*})dx = M^*G,$$

which means that the aggregate land quasi-rent is equal to the cost of the equilibrium number of facilities.

Thus, we have the following:

Proposition 5.8 *The voting equilibrium under a perfectly competitive land market is efficient. Furthermore, the efficient number of facilities is financed solely by the aggregate land quasi-rent.*

This result is reminiscent of Proposition 4.5 derived in a spatial competition model. As observed by Cremer et al. (1983), the same outcome could be reached by applying a spatial tax that is given by

$$t(x) = \frac{1}{Y}\left[\frac{MG}{l} + \frac{l}{M} - \min_{i=1,\ldots,M} t|x - y_i|\right].$$

One way to implement this tax is to subsidize transportation fully while levying a lump sum tax equal to $(M^oG/l) + (tl/4M^o)$.

5.3.4 Notes on the Literature

The process of competition between cities considered in Section 5.2 bears strong resemblances to the one analyzed in Section 4.3. Not surprisingly, therefore, the results are comparable and references overlap. These processes are, however, often cast within different frameworks. Indeed, despite the conceptual similarity between private and public goods in the spatial context, economists studying these subjects have made different modeling assumptions that have different intellectual origins. The literature on local public goods is huge and cannot be reviewed in this book. Distinct and complementary surveys may be found in Wildasin (1986b; 1987) and Scotchmer (1994; 2001). Note that the role of the Henry George theorem in local public finance has been studied independently by Flatters, Henderson, and Mieszkowsi (1974) as well as by Stiglitz (1977) and Arnott and Stiglitz (1979).

A large body of literature also exists devoted to the location of public facilities that is more in the spirit of the material presented in Section 5.3. The aim of these models is to help the decision maker by giving him relevant information about the desirable configurations. Progress in computer science and operations research allows one to solve problems whose size exceeded the capabilities of technology only 10 years ago, and much more progress can be expected in the future. Facility location analysis is a fast-growing domain: the bibliography of Domschke and Drexl (1985) contains over 1,500 entries. Today, there would be many more. Despite significant differences, facility location and local public goods are connected fields; both strands of literature are compared in Thisse and Zoller (1983).

5.4 CONCLUDING REMARKS

The analysis conducted in this chapter leads to fairly strong conclusions. Space blurs the distinction between public and private goods. Together with Chapter 4, this chapter suggests indeed that the supply of local public goods obeys principles that are not fundamentally different from those governing the efficient supply of private goods (Scotchmer and Thisse 1999). In both cases, the working of a perfectly competitive land market seems to be able to improve the allocation of resources vastly in situations that are often described as typical market failures – a property that is overlooked by many in the economics profession. The underlying principle is simple: The differential land rent capitalizes the costs and advantages associated with a particular location, thus fostering the equalization of utility across similar individuals located at different places. Indeed, when the land market is perfectly competitive, the optimal system of cities is identical to the one emerging from competition among land developers. In the same spirit, when there are several public goods, the relevant decision-making entities should be consolidated and incorporated into areas sufficiently large to allow them to internalize the effects of local public policies as much as possible. Finally, even under a political process such as voting, accounting for the differential land rent allows for a reconciliation between the voting outcome and the optimum.

Such results are provocative enough for the problem of land property rights to receive more attention than it does nowadays. Despite the limitations discussed in Section 5.2.5, as well as the imperfections that characterize the housing markets in the real world (Arnott 1995), we may safely conclude that competition among land developers, local governments, or both in the presence of competitive land markets will significantly contribute to the efficient provision of local public goods.[17] In the absence of better alternative mechanisms, land capitalization is worth serious consideration. At the very least, antitrust authorities should be invited to watch more carefully how land and housing markets and developers operate.

It is often forgotten that major debates about land property rights have arisen in the past precisely for the reasons discussed in this chapter. Considering the situation in Europe during the second half of the nineteenth century, Hohenberg and Lee (1985, 326) came to the following conclusion:

> It was recognized that public purposes might require forced purchase of land, for example, for roads . . . But only such land as would actually be used could be appropriated by the collectivity, and any regulation imposed on landowners must be the same for all, whatever their location in the city. Thus large-scale public projects could not recoup their cost by capturing the gains in land value they generated.[18]

It is also worth stressing that the results derived in Section 5.3 suggest a stylized history of the formation of nations in Western Europe. The process of amalgamation of regions started under the ancient regime. As democracy developed in the nineteenth century, centralization of government services was pursued. This is likely because, at the same time, nationalism took a firm hold among the population while substantial technological progress developed in transportation. Indeed, as observed by Cremer et al. (1983) as well as by Alesina and Spolaore (1997), the rise in nationalism can be viewed in the present model as a decline in the parameter t that intensified the decrease generated by technological innovations. Ultimately, a large group of individuals came to favor a geographical concentration of government services. In other words, this was the time of the nation-state.

Around the middle of the twentieth century, a resurgence of regionalism occurred. In our setting, this means an increase in the parameter t. Not surprisingly, geographical decentralization was the answer of most national governments. Our analysis then suggests that the regional system chosen by the majority of the population has been inefficient, thus generating endless debates with those defending centralization of government activities.

As suggested by Alesina and Spolaore (1997), through an increased reliance on voting, the number of regions and countries may increase as the degree of economic integration and openness rises. Indeed, the benefits of a large home market become relatively less important if small countries can freely trade with each other as well as with the rest of the world. Consequently, in the twenty-first century, regional separatism could well be associated with increasing economic integration (Alesina and Spolaore 1997).

APPENDIX

Without loss of generality, we may first divide (5.25) by t and then replace G/nt by G in the corresponding expression.

A. Consider M facilities and let $y_i^* = (2i-1)l/2M$ for $i = 1, \ldots, M$ and $A_1, \ldots, A_i, \ldots, A_M$ be the intervals given by

$$A_1 = [0, l/M[, \ldots, A_i = [(i-1)l/M, il/M[, \ldots, A_M = [(M-1)l/M, l].$$

Consider any configuration with M facilities $\mathbf{y}_M \neq \mathbf{y}_M^*$ (without loss of generality, we assume that all the components of $\mathbf{y}_M$ are distinct) and let N be the number of locations such that $y_i = y_i^*$ with $N < M$. Finally, denote by $I(\mathbf{y}_M^*, \mathbf{y}_M)$ the set of consumers indifferent between the two configurations $\mathbf{y}_M^*$ and $\mathbf{y}_M$ and by $\bar{\mu}$ the measure of the set of consumers who strictly prefer $\mathbf{y}_M$ to $\mathbf{y}_M^*$ (μ is the Lebesgue measure defined on $[0, l]$).

For any j such that $y_j = y_j^*$, we have

$$\mu\{A_j \cap I(\mathbf{y}_M^*, \mathbf{y}_M)\} \leq \frac{l}{M}$$

so that

$$\mu\{I(\mathbf{y}_M^*, \mathbf{y}_M)\} \leq \frac{Nl}{M}. \tag{A.1}$$

Let $B_1, \ldots, B_i, \ldots, B_{M+1}$ be the intervals of $[0, l]$ defined as follows:

$$B_1 = [0, l/2M[, \ldots, B_i = [(2i-3)l/2M, (2i-1)l/2M[, \ldots, B_{M+1}$$
$$= [(2M-1)l/2M, l].$$

Denote by k_i the number of facilities in $\mathbf{y}_M$ belonging to B_i and by μ_i the measure of the set of consumers in B_i who strictly prefer $\mathbf{y}_M$ to $\mathbf{y}_M^*$. For $i = 2, \ldots, M$, it can readily be verified that $\mu_i = 0$ if $k_i = 0$, $\mu_i = l/2M$ if $k_i = 1$, and $\mu_i < l/M$ if $k_i \geq 2$. Similarly, for $i = 1$ and $i = M+1$, we have $\mu_i = 0$ if $k_i = 0$ and $\mu_i < l/2M$ if $k_i \geq 1$, respectively. Let $\hat{M} < M$ be the number of intervals B_i for which $k_i \geq 2$. The following three cases may then arise:

1. $k_1 = k_{M+1} = 0$. Then, there are at most $(M - 2\hat{M} - N)$ intervals B_i containing one facility and

$$\bar{\mu} \leq \hat{M}\frac{M}{l} + (M - 2\hat{M} - N)\frac{l}{2M},$$

the inequality being strict when $\hat{M} > 0$.

2. $k_1 = 0$ and $k_{M+1} \geq 1$ (or, symmetrically, $k_1 \geq 1$ and $k_{M+1} = 0$). Then, we have

$$\bar{\mu} \leq \frac{l}{2M} + \hat{M}\frac{l}{M} + (M - 1 - 2\hat{M} - N)\frac{l}{2M}$$

3. $k_1 \geq 1$ and $k_{M+1} \geq 1$. Then, we have

$$\bar{\mu} \leq \frac{l}{M} + \hat{M}\frac{l}{M} + (M - 2 - 2\hat{M} - N)\frac{l}{2M}.$$

In all cases, it follows that

$$\bar{\mu} \leq (M - N)\frac{l}{2M}. \tag{A.2}$$

Assume that the inequality is strict in (A.2). Then, (A.1) and (A.2) imply that μ^*, the measure of the set of consumers who strictly prefer $\mathbf{y}_M^*$ to $\mathbf{y}_M$, is strictly larger than $\bar{\mu}$, thus implying that $\mathbf{y}_M$ is defeated by $\mathbf{y}_M^*$.

Suppose now that the equality holds in (A.2). Hence, it follows that $k_1 = k_{M+1} = 0$ and $\hat{M} = 0$. Consequently, an interval B_j exists containing one facility and having one facility placed at one of its endpoints such that either

$$\mu\{A_{j-1} \cap I(\mathbf{y}_M^*, \mathbf{y}_M)\} < l/M$$

or

$$\mu\{A_j \cap I(\boldsymbol{y}_M^*, \boldsymbol{y}_M)\} < l/M$$

holds. This implies that $\mu^* > \bar{\mu}$ so that $\boldsymbol{y}_M$ is again defeated by $\boldsymbol{y}_M^*$.

B. When there are M facilities, total transportation costs are as follows:

$$\begin{aligned} TTC(y_1, \ldots, y_M) &= \int_0^{(y_1+y_2)/2} |x - y_1| dx \\ &\quad + \sum_{i=2}^{M-1} \int_{(y_{i-1}+y_i)/2}^{(y_i+y_{i+1})/2} |x - y_i| dx \\ &\quad + \int_{(y_{M-1}+y_M)/2}^{l} |x - y_M| dx \\ &= \frac{y_1^2}{2} + \sum_{i=1}^{M-1} \frac{(y_{i+1} - y_i)^2}{4} + \frac{(l - y_M)^2}{2}. \end{aligned}$$

Applying the first-order conditions to this expression yields the linear system

$$3y_1 - y_2 = 0$$

$$-y_{i-1} + 2y_i - y_{i+1} = 0 \qquad \text{for } i = 2, \ldots, M - 1$$

$$-y_{M-1} + 3y_M = 2l,$$

which has a single solution given by $y = (2i - 1)l/2M$. That the second-order conditions are always satisfied can easily be checked.

C. First, some simple calculations show that $\hat{M}$ defeats $M < \hat{M}$ when $\hat{M} \leq 4$. It then remains to prove that $\hat{M}$ defeats any $M > \hat{M}$. Let $\hat{\mu}$ be the measure of the set of consumers who strictly prefer $\hat{M}$ to $M > \hat{M}$ and show that $\hat{\mu} > l/2$.

Assume first that $M > (3\hat{M} + 1)/2$. For each $i = 1, \ldots, \hat{M}$, a consumer located at

$$x \in \left[\frac{2i-1}{M}\frac{l}{2} - \frac{l}{2M} \min\left\{ \frac{M - \hat{M}}{M + 1}, 1 \right\}, \frac{2i-1}{M}\frac{l}{2} + \frac{l}{2M} \min\left\{ \frac{M - \hat{M}}{\hat{M} + 1}, 1 \right\} \right]$$

strictly prefers $\hat{M}$ to M because

$$\min_{i=1,\ldots,\hat{M}} |x - y_i^*(\hat{M})| - \min_{i=1,\ldots,M} |x - y_i^*(M)| < \frac{(M - \hat{M})l}{2(\hat{M} + 1)\hat{M}},$$

where

$$y_i^*(M) = (2i - 1)l/2M \qquad i = 1, \ldots, M.$$

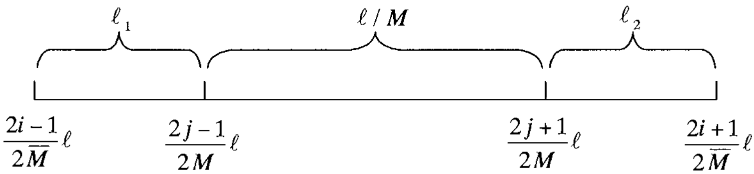


Figure 5.2: Population segmentation.

Therefore, $\hat{\mu}$ is larger than or equal to

$$\min\left\{\frac{M-\hat{M}}{\hat{M}+1}, 1\right\} l$$

so that $\hat{\mu} > l/2$ since $M > (3\hat{M}+1)/2$.

Assume now that $\hat{M} < M \leq (3\hat{M}+1)/2$. To start with, consider the interval $[0, l/2M]$. We know that $l/2M < l/2\hat{M} < 3l/M$. Consequently, the consumers in $[0, l/2M]$ strictly prefer $\hat{M}$ to M if and only if $1/M\hat{M} < 2G/l^2$ – a condition that is satisfied by definition of $\hat{M}$. The same holds for the interval $[(2\hat{M}-1)l/2\hat{M}, l]$. Consider now an interval given by

$$C_i =](2i-1)l/2\hat{M}, (2i+1)l/2\hat{M}[$$

for $i = 2, \ldots, M-1$. Two subcases may arise. In the first one, there is one facility in the equidistant configuration with M facilities, which belongs to C_i. Then, the measure of the set of consumers who strictly prefer $\hat{M}$ to M is larger than $l/2\hat{M}$. In the second one, there are two facilities of that configuration that belong to C_i, as represented in Figure 5.2.

1. If both l_1 and l_2 are smaller than $(G/l)(M-\hat{M})$, then the consumers at locations $(2j-1)l/2M$ and $(2j+1)l/2M$, and, therefore, all consumers in C_i, strictly prefer $\hat{M}$ to M.
2. If only l_1, say, is smaller than $(G/l)(M-\hat{M})$, the consumers located in

$$\left]\frac{(2i-1)l}{2\hat{M}}, \frac{(4j-1)l}{2M}\right]$$

and in

$$\left]\frac{(2j+1)l}{4M} + \frac{(2i+1)l}{4M}, \frac{(2i+1)l}{2\hat{M}}\right]$$

strictly prefer $\hat{M}$ to M.
3. Finally, observe that both l_1 and l_2 cannot exceed $(G/l)(M-\hat{M})$ because

$$l_1 + l_2 = \frac{l}{\hat{M}} - \frac{l}{M} < \frac{(M-\hat{M})l}{(\hat{M}+1)\hat{M}} < \frac{2G}{l}(M-\hat{M})$$

since $M > \hat{M}$. Thus, in the three cases, the measure of the set of consumers in C_i who strictly prefer $\hat{M}$ to M is larger than $l/2\hat{M}$.

Summing over all the intervals, we obtain

$$\hat{\mu} > \frac{l}{\hat{M}} + (M-1)\frac{l}{2\hat{M}} > l/2.$$

D. It is sufficient to show that M defeats $M-1$ as long as $M \leq \hat{M}$. This, together with Part **C**, implies that $\hat{M}$ is the unique value of M that simultaneously defeats $M-1$ and $M+1$.

The consumers who strictly prefer M to $M-1$ are located either in the hinterlands or around the M facilities of the equidistant configuration. Consider the first group. The consumer at $l/2M$ strictly prefers M to $M-1$ because of (5.11). This holds a fortiori for all the consumers in $[0, l/2M]$. The same applies to consumers in $[(2M-1)l/2M, l]$. We now consider the intermediate groups. Specifically, we focus on the consumers located near the facility established at $(2i-1)l/2M$ for $i = 2, \ldots, M-1$. Clearly, we have

$$\frac{2i-3}{M-1} < \frac{2i-1}{M} < \frac{2i-1}{M-1}.$$

Let x_i^l be the location of the consumer situated on the left-hand side of the ith facility and indifferent between M and $M-1$. Because x_i^l must belong to the interval

$$\left]\frac{(2i-3)l}{2(M-1)}, \frac{(2i-1)l}{2M}\right]$$

it follows that x_i^l is the solution to

$$x_i^l - \frac{2i-3}{2(M-1)}l - \left(\frac{2i-1}{2M}l - x_i^l\right) = \frac{G}{l},$$

that is,

$$x_i^l = \frac{2i-3}{4(M-1)}l + \frac{2i-1}{4M}l + \frac{G}{2l}.$$

Similarly, if x_i^r is the location of the consumer situated on the right-hand side of the ith facility and indifferent between M and $M-1$, we obtain

$$x_i^r = \frac{2i-1}{4(M-1)}l + \frac{2i-1}{4M}l - \frac{G}{2l}.$$

All consumers in $]x_i^l, x_i^r[$ strictly prefer M to $M-1$. Denote by μ_M^* the measure of the set of consumers who strictly prefer M to $M-1$. It then follows from the foregoing and from the continuity of the utility functions that

$$\mu_M^* > \frac{l}{M} + (M-2)\left[\frac{l}{2(M-1)} - \frac{G}{l}\right].$$

Because $M \leq \hat{M}$, we have

$$\frac{l}{2(M-1)M} > \frac{G}{l}$$

so that $\mu_M^* > l/2$.

NOTES

1. For example, the German emperor Henry V affirmed the principle "City air brings freedom" in charters for Speyer and Worms.
2. Other examples of local public goods that have played a major role in the history of cities are religious temples, royal palaces, and public area places such as the agora or the forum.
3. The distinction between *traveled-for goods* and *delivered goods* made by Lea (1979) is not essential for our purpose. It is indeed reasonable to believe that the quality of a delivered good often decreases as the distance increases. So, in both cases, users' benefits are subject to a distant-decay effect.
4. Another interpretation is to consider a monocentric city of Section 3.3. In this context, the following analysis remains valid if $T(r)$ is interpreted as the commuting cost to the CBD, whereas the local public good is uniformly available within the entire city.
5. A brief overview of the work of George is provided by Whitaker (1998).
6. This question has generated harsh political debates in the United States. See Mills (1972b, chap. 3) for a critical appraisal of the welfare and ethical aspects of the single tax.
7. In a planning approach without a land market, we should account for the land availability constraint at each *r*. The corresponding multipliers would then correspond to the equilibrium land rent used here.
8. If this optimization problem has several solutions, several optimal urban systems exist. In this case, the following discussion applies to each one of them.
9. Formally, μ is the Lebesgue measure.
10. Using the notation introduced in the preceding section, we also have $N^o = N^o(G^o)$.
11. To study existing communities with fixed boundaries, there is an alternative approach to the land developer model in urban public finance (Wildasin 1986b; 1987). Residents are voters who, according to the median voter rule, choose the quantity of public goods and the taxation scheme to be established within the community. In this case, consumers may be heterogeneous in terms of income, tastes, or both. This leads to a segmentation of the population as envisioned by Tiebout, whereas consumers are homogenous in the foregoing analysis. A model in that spirit will be considered in the next section.
12. In a sense, this problem is similar to the one with neighborhood public goods, the difference being that here a priori there is no center.

13. As an example, think of the many difficulties faced by public authorities when they try to implement urban tolls or peak-load pricing within cities.
14. Alternatively, we may interpret the model as one in which individuals vote on the number of regions and the location of capitals.
15. Because Y is fixed, this is equivalent to the lump sum tax scheme used in the previous section.
16. By choosing $\theta(x)$ appropriately, we may guarantee that (5.33) always holds.
17. See, however, Henderson and Thisse (2001) as well as Helsley and Strange (1997) for additional limitations of the land development process when strategic considerations are taken into account.
18. For example, Hausmann has financed property acquisition and construction against future revenues obtained from the increased property values created by the planned improvements of Paris (Barnett 1986; Marchand 1993). Interestingly, the operation went bankrupt when the landowners recouped their property at the prices prevailing before the improvements (Marchand 1993).

PART II

THE STRUCTURE OF METROPOLITAN AREAS

6

The Spatial Structure of Cities under Communications Externalities

6.1 INTRODUCTION

The most distinctive feature of a city is its much higher population density than the surrounding nonurban areas. As a result, economic agents residing within a city are close to one another. But why do households and firms seek spatial proximity? Fundamentally, this occurs because economic agents need to interact and distance is an impediment to interaction. This need is gravitational in that its intensity is likely to increase with the number of agents set up in each location and to decrease with the distance between two locations. This need has been at the heart of the work of several geographers, and we will encounter it on many occasions and with different economic meanings.

However, by crowding a few locations only, economic agents also decrease their satisfaction because they normally enjoy consuming more land, either as consumers or as producers. Therefore, *one can view the agglomeration process*, at least in the first order, *as the interplay between an interaction field among agents and competition on the land market*. In such a setting, the need to interact acts as a centripetal force, whereas competition for land has the nature of a centrifugal force. As will be seen in this chapter, it is remarkable that the mere need to interact turns out to be sufficient to generate a single-peaked distribution of (homogeneous) agents across locations.

But this is not the end of the story. Indeed, one has to explain why economic agents want to interact? It should be clear that several explanations can be put forward. That human beings are "social animals" is perhaps the most basic justification of the need for interaction among individuals. Indeed personal relations are the essence of societies even though the consequences of relations are often double-edged. For example, according to Fisher (1982, 2–3),

> Our day-to-day lives are preoccupied with people, with seeking approval, providing affection, exchanging gossips, falling in love, soliciting advice, giving opinions, soothing anger, teaching manners, providing aids, making impressions, keeping in touch. . . . Although modern nations have elaborate arrays of institutions and organizations, daily

life proceeds through personal ties. . . . Those personal ties are also our greatest motives for action.

Psychologists also recognize that human beings have a pervasive drive to form and maintain lasting and positive relations with others. According to Baumeiter and Leary (1995, 497), who reviewed a vast literature, two conditions must be met for this drive to be satisfied:

First, there is a need for frequent, affectively pleasant interactions with a few other people, and, second, these interactions must take place in the context of a temporally stable and enduring framework of affective concern for each other's welfare.

To the best of our knowledge, the first economic model focusing on such a trade-off was proposed by Beckmann (1976). More precisely, Beckmann assumes that the utility of an individual depends on the average distance to all individuals with whom this person interacts as well as on the amount of land she buys on the market. Under such preferences, the city exhibits a bell-shaped population density distribution supported by a similarly shaped land rent curve. Thus, the city emerges here as a social magnet. Stated differently, the natural gregariousness of human beings leads to the spatial concentration of people within compact areas. This model is presented in Section 6.2.1.

Although the process of interaction goes both ways, individuals worry only about their role as "receivers" and tend to neglect their function as "transmitters" to others. Hence, the equilibrium distribution of agents within the city is unlikely to be an optimum. Indeed, a comparison of the equilibrium and optimum densities shows that the former is less concentrated than the latter. This suggests, from the social standpoint, that the need to interact may well result in an insufficient concentration of population around the city center. Contrary to general beliefs, therefore, it is not obvious that agents are too densely packed in cities.

At this stage, it is natural to ask whether the principles uncovered by Beckmann for households also govern the locational decisions made by firms within an urban area. In raising such a question, one may wonder what is the nature of the interaction that would foster firms' concentration beyond the standard market transactions in which they are involved. The reason here is very different from the one we saw for consumers in that it refers to the role of information as a basic input in firms' activities. By this, we mean the kind of information that is difficult to codify because it is tacit, and thus it can typically be collected only through face-to-face communications that require travel by high-skilled people whose time is valuable.

The impact of information on locational decisions is not new. The decline of manufacturing employment and the growth of office employment in central cities has been a common trend observed in many countries that started after the Industrial Revolution.[1] For example, in their study of the urban making of Europe, Hohenberg and Lee (1985, 299) forcefully argue that the common element of the tertiary or service activities of cities is information, an intangible and therefore bulkless commodity that manifests itself mainly in the act of being transferred or exchanged. Town centers were the natural location where those trafficking in knowledge congregated, and they displaced not only residents but also most activities dealing with visible commodities. The business center was taken over by an army of brokers, clerks, bankers, couriers, and other dealers in the quintessentially urban commodity, information.

A fundamental characteristic of information is its public-good nature: the use of a piece of information by a firm does not reduce the content of that information for other firms. Hence, the exchange of information through communication within a set of firms generates externality-like benefits for each of them (Stigler 1961). Provided that firms own different types of information, the benefits of communication generally increase as the number of firms rises. The quality of the information is also better when firms are gathered in that the number of intermediates is smaller. Because communications typically involve distance-decay effects, a fact well documented since the pioneering work of Hägerstrand (1953), the benefits are greater if firms locate closer to each other.

The empirical evidence is fairly conclusive. Controlling for the geographical concentration of sectors affecting the location of patent use, Jaffe, Trajtenberg, and Henderson (1993) found that, in the United States, citations of patents are more likely to be domestic and to come from the same states and metropolitan statistical areas, thus suggesting that the diffusion of knowledge is spatially concentrated (at least at the early stages of the diffusion process). This conclusion is strengthened by several other empirical studies, such as the one of Audretsch and Feldman (1996), who observed that external spillovers are likely to be geographically bounded within the region where the new knowledge was created.

In this respect, it is well known that face-to-face communication is most effective for rapid product and process development when the access to information about new products and production processes turns out to be essential for the competitiveness of firms. Most likely, the origin of these spillovers lies in the existence of face-to-face contacts. And indeed, in their survey of empirical evidence, Tauchen and Witte (1984) observed that much of the interaction among the employees of different firms consists of such contacts. For example, Saxenian (1994, 33) emphasized the importance of this factor in making the Silicon Valley an efficient productive system:

> By all accounts, these informal conversations were pervasive and served as an important source of up-to-date information about competitors, customers, markets, and technologies. Entrepreneurs came to see social relationships and even gossips as a crucial aspect of their business. In an industry characterized by rapid technological change and intense competition, such informal communication was often of more value than more conventional but less timely forums such as industry journals.

The key point here is that personal contacts within the agglomeration encourage a constant intercommunication of ideas. This might come as a surprise in the age for which futurists had predicted the decline of cities because people would use more and more telecommunications devices instead of face-to-face interaction (see, e.g., Toffler 1980). However, it is well known that a substantial amount of knowledge used by firms turns out to be tacit and difficult to transfer from one location to another (see, e.g., Teece 1977 for an early contribution). The difference between *tacit* and *codified* information (knowledge) is crucial here. The transfer of information through modern transmission devices requires its organization according to some prespecified patterns, and only formal information can be codified and sent to others in this way. For example, the initial steps in the development of a new technology require repeated contacts between the actors involved to develop a mutual way of communicating through some common codes, to figure out how to interpret personalized information, and to make them operational. Such a process is facilitated by spatial proximity.

Furthermore, the historical evidence regarding the impact of the telephone on urbanization suggests a positive correlation between city size and telephone use (Gaspar and Glaeser 1998). Although telecommunications may be a substitute for face-to-face meetings, these two forms of communications may also be complementary. For example, Gaspar and Glaeser (1998) reported on some suggestive evidence that an increase in business trips has occurred despite (or because of) recent improvements in telecommunications technologies. Thus, contrary to the opinions of futurists, the development of such technologies does not (necessarily) imply the death of cities as information centers.

In essence, the explanation is that the transmission of knowledge and ideas is not a routine activity that can be performed through standardized procedures. It is a cognitive process (and uncertainty is therefore inherent to the exchange) that is made easier when the individuals involved are close to each other. As Feldman (1994, 2) put it nicely, "knowledge traverses corridors and streets more easily than continents and oceans." Furthermore, face-to-face communications are often at the origin of new ideas, combining insights from each party that are crucial for innovations.

In two independent articles, Borukhov and Hochman (1977) and O'Hara (1977) have shown how face-to-face communications may induce (office) firms to congregate and to form a central business district, even though clustering results in higher land rents. A similar problem has been studied recently by Lucas (2001) in a more general context. This topic is discussed in Section 6.2.2.

Diversity is another fundamental distinctive feature of cities. In other words, cities are concentrations of different agents (mainly firms and households). The mere recognition of this simple fact should lead to a new and richer set of results. The next step is, then, to mix consumers and firms within broader models in order to study what the interplay between the two groups of agents may result in. The centripetal force in this interplay is the communications among firms permitting the exchange of information: other things being equal, each firm has an incentive to establish itself near other firms, thus fostering agglomeration. The centrifugal force is less straightforward and goes through the land and labor markets. The clustering of many firms in a single area increases the average commuting distance for their workers, which, in turn, increases the wage rate and land rent in the area surrounding the cluster. Such high wages and land rents tend to discourage further agglomeration of firms in the same area. The equilibrium distribution of firms and households is thus the balance between these opposite forces. In Section 6.3, we focus on the direct interaction between firms because we believe that they are even more fundamental in modern societies for the shaping of cities than social interactions among individuals.

The interplay between those two types of forces has been studied in a series of articles by Fujita, Imai, and Ogawa. These authors have shown that different equilibrium patterns may emerge according to the values of the economy's basic parameters, thus affecting the balance of the two opposite forces. More surprising, perhaps, is that the shape of the interaction field also influences the types of equilibria that arise. We consider in Section 6.4 a linear accessibility field and show that the market solution is essentially monocentric. When commuting costs are high in relation to the accessibility parameter measuring the importance of the distance-decay effect in the interaction field, the equilibrium involves a complete mixture of business and residential activities. As the commuting cost falls, two business districts, which are themselves flanked by a residential area, are formed around the integrated section that shrinks. Eventually, *when commuting costs are low enough, the city becomes monocentric with the emergence of a single business district surrounded by two residential sections*. This seems to accord with what we have observed since the beginning of the technological revolution in transportation.

Furthermore, the monocentric configuration is socially desirable for a larger domain of parameters than for the equilibrium outcome. In other words, the market may lead to a more dispersed configuration of firms than would be socially optimal.

As examined in Section 6.5, the possible equilibrium patterns are much more complex in the case of an exponential distance-decay function. In addition to the configurations just mentioned, we will see that intermediate values for the commuting costs may lead to a duocentric configuration, or to a configuration involving a primary business center and two secondary business centers, or to three more or less identical centers. In these last three cases, the equilibrium pattern may also be viewed as describing a system of two or three cities.

Even more interesting is the nature of the transition from one equilibrium to another when some parameters slightly change. For example, one may observe catastrophic modifications in the urban configuration when the equilibrium city moves from a monocentric configuration to a duocentric one in which the interior residential area is quite large although the distance-decay parameter changes only slightly. All these results show how nonlinearities in accessibility may lead to a vast set of different outcomes and, by the same token, may explain why it is often hard to make reliable predictions about urban development.

Finally, in Section 6.6, we briefly discuss the case in which a firm splits its activities between two units located far apart. In this way, we may capture the idea that some activities are crucially dependent on the information obtained from other firms, whereas more routine activities require communication with the firm's headquarters only. The typical configuration associated with low intra-firm communication costs is the agglomeration of front offices at the city center and the dispersion of back offices together with their workers in the suburbs.

In all models studied in this chapter, a city emerges in an otherwise homogeneous space as the collective outcome of the interplay between individual decision makers. Cities are not the result of the actions taken by land developers or local governments. Furthermore, we focus on the formation and spatial structure of a city but do not address the issue of its size. Thus, all the results derived in this chapter are to be understood relative to given populations of households and firms.

Before proceeding, we want to stress that, although we work with a continuum of agents, we assume that each agent is indivisible. The implications of this assumption are important for the models examined in this chapter as well as in subsequent ones because the results are partly determined by the indivisibility of firms and households over several locations (even in the case of the multiunit firms, units cannot be subdivided). In other words, agents have a well-defined spatial identity.[2] The advantages of such a modeling strategy were made clear by Hotelling (1929) but became broadly accepted in the economics profession only after the work of Aumann (1964).[3]

6.2 AGGLOMERATION AS SPATIAL INTERACTION AMONG INDIVIDUALS OR FIRMS

6.2.1 Social Externalities between Individuals

The propensity to interact with others is a fundamental human attribute. People like to be close to each other to maximize social interaction. The following simple model demonstrates how the preference for social life leads to the emergence of a center through a unimodal and symmetric distribution of individuals. This distribution is dispersed around the center because competition for land leads to higher land rents near the center in a market economy.[4]

Consider a continuum N of identical consumers and a one-dimensional space $X = (-\infty, \infty)$ in which the land density is 1 everywhere and the opportunity cost of land is $R_A > 0$. Land is owned by absentee landlords. A consumer residing at x is endowed with the preferences

$$U = u(z, s) + I_x,$$

where, as usual, z and s denote the composite good and the lot size, respectively, and I_x stands for the *interaction field* of a consumer located at x whose budget constraint is

$$z + sR(x) = Y - T(x).$$

Here $T(x)$ is the associated travel cost borne by the consumer. Because trips are costly, the occupied area must be bounded. Without loss of generality, the area under consideration is denoted by $[-b, b]$, where b is the agglomeration boundary determined at the equilibrium. As in Section 3.2, the quantity of the composite good is obtained by solving $u(z, s) = U - I_x$ for z, which is represented by $z = Z(s, U - I_x)$; thus, the consumer's bid rent function is given by

$$\Psi(x, U) = \max_s \frac{Y - Z(s, U - I_x) - T(x)}{s}. \tag{6.1}$$

In analyzing this model, Beckmann (1976) made several simplifying assumptions to derive a closed-form solution. First, the utility u is expressed as

$$u(z, s) = z + \alpha \log s$$

for which $\alpha > 0$ is a parameter indicating the weight of land in consumer preferences. Second, the interaction field is supposed to be such that each consumer interacts once with each and every other consumer in the area under consideration, which represents the extreme view that the individual network of personal ties is formed by the whole population residing in the urban area under consideration. Hence, the utility of I_x is constant across locations

$$I_x = I.$$

At best, this assumption is to be considered as a first exploration of one of the main determinants of a city existence. Finally, maintaining a bond requires time, money, and attention. It is assumed that this takes the form of a trip whose cost is linear in distance, and thus the total travel cost borne by a consumer at x while interacting with others is given by

$$T(x) \equiv \int_{-b}^{b} t|x - y|n(y)dy, \tag{6.2}$$

where $n(y)$ is the density of consumers at location y and $t > 0$ the unit travel cost. It is worth noting that, in $T(x)$, the travel cost from x to y is weighted by the number (formally the density) of individuals located at the destination y, which means that there is one trip per individual at y. Clearly we have

$$T(x) = \int_{-b}^{x} t(x - y)n(y)dy + \int_{x}^{b} t(y - x)n(y)dy, \tag{6.3}$$

and thus $T(x)$ varies with the consumer's location as well as with the entire population density. It should be clear that the cost borne by a consumer when interacting with others is here very different from what it was in Chapter 3, where all activities were supposed to take place at the city center.

Because all consumers must reach the same utility level U^* in equilibrium, (6.1) may be rewritten as

$$\Psi(x, U^*) = \max_s \frac{Y - U^* + I + \alpha \log s - T(x)}{s}. \tag{6.4}$$

The first-order condition with respect to s gives the following equilibrium condition:

$$Y - U^* + I - \alpha + \alpha \log s - T(x) = 0. \tag{6.5}$$

Let

$$\zeta \equiv Y - U^* + I - \alpha,$$

where ζ is a constant whose value is unknown. Solving (6.5) for s and using the definition of ζ, we obtain

$$s^*(x) = \exp\left(\frac{-\zeta + T(x)}{\alpha}\right) \tag{6.6}$$

so that

$$n^*(x) = \exp\left(\frac{\zeta - T(x)}{\alpha}\right), \tag{6.7}$$

where $n^*(x) \equiv 1/s^*(x)$ is the equilibrium population density. Substituting (6.6) into (6.4) leads to

$$\Psi(x, U^*) = \frac{\alpha}{s^*(x)}. \tag{6.8}$$

Because $\Psi(x, U^*) = R_A$ at the equilibrium city fringe b^*, we must have

$$\frac{R_A}{\alpha} = n^*(b^*) = \exp\left(\frac{\zeta - T(b^*)}{\alpha}\right). \tag{6.9}$$

Differentiating (6.3) twice yields

$$\frac{d^2T}{dx^2} = 2tn^*(x),$$

and thus $T(x)$ is strictly convex in x. It then follows from (6.7) that

$$\frac{d^2T}{dx^2} = 2t \ \exp\left(\frac{\zeta - T(x)}{\alpha}\right). \tag{6.10}$$

Solving this differential equation yields

$$T(x) = -\alpha \ \log\left[\frac{\alpha}{t} \ \exp\left(-\frac{\zeta}{\alpha}\right)\frac{k^2 \ \exp(k|x|)}{(1 + \exp(k|x|))^2}\right], \tag{6.11}$$

where k is an unknown positive constant of integration (see Part A of the chapter appendix for detailed calculations). Substituting (6.11) into (6.7), we obtain

$$n^*(x) = \frac{\alpha}{t}\frac{k^2 \exp(k|x|)}{(1+\exp(k|x|))^2}, \tag{6.12}$$

which gives the equilibrium population density as a function of the constant k.

Clearly this function is symmetric about the origin. Furthermore, differentiating (6.12) with respect to x shows that $n^*(x)$ has a unique maximum at $x = 0$. The equilibrium consumer density is therefore unimodal. Taking the second derivative shows that $n^*(x)$ is concave over the interval

$$\left[-\frac{\log(2+\sqrt{3})}{k}, \frac{\log(2+\sqrt{3})}{k}\right]$$

and convex outside. A typical pattern of the population density is depicted in Figure 6.1.

Accordingly, a preference for social life is sufficient to form an agglomeration around an endogenous center and, therefore, to preclude the flat distribution of individuals from being an equilibrium. This center is the place where the interaction among people is most convenient and individual land consumption is the lowest. And, indeed, the population is more concentrated around the center because the population density falls when the distance from the center rises, thus confirming the intuition that being at the center of the urban area gives the individuals located there the highest accessibility to others. This implies that the center emerging here plays the same role as the CBD in the urban economics models surveyed in Chapter 3. The crucial difference is that the center is endogenous here, whereas the CBD considered in Chapter 3 is exogenous.

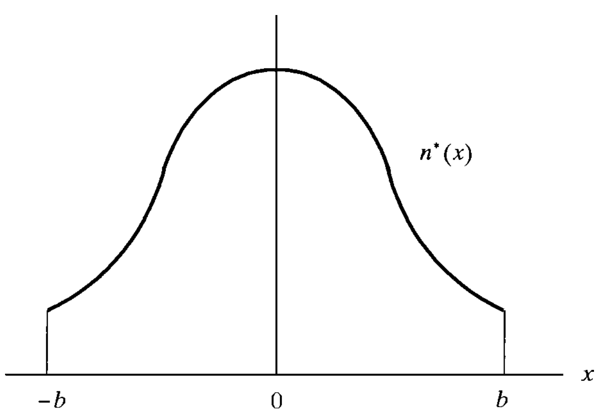


Figure 6.1: The equilibrium population density when consumers like to interact.

Using (6.9) and (6.12), we see that the following condition must hold at the urban fringe:

$$\frac{R_A}{\alpha} = \frac{\alpha}{t} \frac{k^2 \exp(kb)}{(1 + \exp(kb))^2}. \tag{6.13}$$

Furthermore, integrating the population density over the interval $[-b, b]$ must be equal to N:

$$\begin{aligned} N &= 2 \int_0^b n^*(x)dx = 2\frac{\alpha}{t} \int_0^b \frac{k^2 \exp(kx)}{(1 + \exp(kx))^2} dx \\ &= 2\frac{\alpha}{t} k \frac{\exp(kb) - 1}{\exp(kb) + 1}. \end{aligned} \tag{6.14}$$

These two expressions give the equilibrium values of the constant of integration k and of the agglomeration boundary b. Setting $y = \exp(kb)$ in (6.13) and (6.14), solving, (6.14) for y, and replacing the value thus obtained in (6.13), we obtain

$$k^2 = \frac{t}{\alpha^2}\left(\frac{tN^2}{4} + 4R_A\right). \tag{6.15}$$

Hence, by (6.7) we have

$$n^*(0) = \frac{1}{\alpha}\left(\frac{tN^2}{16} + R_A\right); \tag{6.16}$$

therefore the peak of the population distribution moves upward (and the population density is more concentrated) when the population size N is larger, when the unit cost of traveling t rises, and when the opportunity land cost R_A increases; the peak of the population distribution moves downward when the preference for land α gets stronger. These results are similar to those derived in Section 3.3.2, but the center is no longer prespecified but determined as the outcome of the interaction between consumers.

Similarly, the urban fringe may be obtained by replacing k in (6.14) with the positive root of (6.15). Using this root then permits the complete determination of the equilibrium population density, which, in turn, is used to find $T(x)$ by (6.3), thus yielding the equilibrium value of ζ by (6.11).

Setting $R^*(x) = \Psi(x, U^*)$ and $1/s^*(x) = n^*(x)$ in (6.8) implies

$$R^*(x) = \alpha n^*(x),$$

and thus the market land rent supporting the residential equilibrium mirrors the population density. Both densities are bell-shaped and vary in the same way with the parameters of the economy. In particular, a rise (fall) in the population (the cost of trips) is reflected in higher (lower) land rents near the urban center. Again, this agrees with what we saw in Chapter 3.[5]

We may summarize these results as follows:

Proposition 6.1 *Assume that consumers value land and interaction with others. Then, if consumers' utility is given by*

$$U = \alpha \log s + I - T(x),$$

the equilibrium population distribution and the equilibrium land rent are unimodal and symmetric.

The indivisibility of human beings is fundamental for this proposition to hold. Indeed, if each individual were to be perfectly divisible, spatial equilibrium would involve a uniform distribution of people and a flat land rent.

We now study the optimal distribution of consumers using the same approach as in Section 3.3.3. Though our equilibrium is competitive, it is not socially optimal. As noted by Tauchen and Witte (1984), the explanation for this market failure is that, although the location of each individual directly affects the travel costs of all others,[6] the consumer considers only own travel costs in making a locational decision. That the interaction among individuals goes both ways and does not transit through the market implies an externality preventing the equilibrium from being at an optimum. However, it is interesting to know whether the equilibrium is excessively or insufficiently concentrated around the center.

To answer this question, we have to provide a fairly detailed analysis of the optimum density. Because they are identical, all the consumers at the same location x must have the same consumption bundle (z, s). Consequently, if U^* denotes the equilibrium utility level obtained above, the optimal distribution must minimize the total cost as defined by (3.26) with respect to $s(x)$ and b:

$$C \equiv \int_{-b}^{b} \{T(x) + Z[s(x), U^*] + R_A s(x)\} n(x) dx,$$

where $Z[s(x), U^*]$ is equal to $U^* - \alpha \log s(x) - I$ and $T(x)$ is given by (6.3) subject to the land constraint

$$s(x)n(x) = 1 \quad \text{for each } -b \le x \le b, \tag{6.17}$$

the population constraint

$$\int_{-b}^{b} n(x) dx = N, \tag{6.18}$$

and the nonnegativity constraints on s and n.

As in Section 3.3.3, one can show that solving this maximization problem is equivalent to solving the following maximization problem:

$$S \equiv NY - C = \int_{-b}^{b} \{[Y - U^* + I + \alpha \log s(x) - T(x)] n(x) - R_A\} dx \tag{6.19}$$

subject to the aforementioned constraints.

We present here an intuitive argument showing how the optimality conditions are obtained.[7] For that, we use (6.17) in order to rewrite our optimization problem as the maximization of the expression

$$S = \int_{-b}^{b} \{[Y - U^* + I - \alpha \log n(x) - T(x)]n(x) - R_A\}dx \tag{6.20}$$

subject to (6.18), that is, we want maximize the Lagrangian function

$$\begin{aligned} L &= \int_{-b}^{b} \{[Y - U^* + I - \alpha \log n(x) - T(x)]n(x) - R_A\}dx \\ &\quad + \lambda\left[\int_{-b}^{b} n(x)dx - N\right] \\ &= \int_{-b}^{b} \{[Y - U^* + I + \lambda - \alpha \log n(x) - T(x)]n(x) - R_A\}dx - \lambda N, \end{aligned} \tag{6.21}$$

where λ is the multiplier associated with the population constraint and U^* is supposed to be given by assumption.

We first choose $n(x)$ at every location x inside the city to maximize L. If we neglect for the moment the impact of the choice of $n(x)$ on $T(x)$, the marginal benefit of increasing the population at x is measured by

$$Y - U^* + I + \lambda - \alpha \log n(x) - T(x) - \alpha.$$

However, by increasing the number of consumers at x by one unit, we increase the travel costs of the other consumers by $T(x)$ owing to the symmetry in transport cost between any two locations, and thus the net benefit generated by one more consumer at x is equal to

$$Y - U^* + I + \lambda - \alpha \log n(x) - T(x) - \alpha - T(x).$$

The optimum is therefore reached when this magnitude equals zero:

$$Y - U^* + I - \alpha + \lambda + \alpha \log s(x) - 2T(x) = 0. \tag{6.22}$$

This condition states that, when the number of consumers at x increases by one, the transport cost at this location increases by the amount $T(x)$. However, the transport cost of the other consumers at x also rises by the same amount. Therefore, in the optimum problem, one must account for both costs in choosing the location of a consumer. By contrast, in the equilibrium problem, in choosing a location at x, a consumer only accounts for her travel cost to the others but neglects the impact of this choice over the travel costs borne by the others (which is also given by $T(x)$).

Setting

$$\zeta^o \equiv -\alpha + Y - U^* + I + \lambda,$$

we obtain from (6.22) the optimal distribution:

$$n^o(x) = 1/s^o(x) = \exp\left(\frac{\zeta^o - 2T(x)}{\alpha}\right), \tag{6.23}$$

which is identical to (6.7) except that the total travel cost $T(x)$ enters with the factor 2 instead of 1.

It remains to determine the optimal city boundary. As above, if we neglect the impact of a marginal increase in b on $T(x)$ and assume symmetry of the distribution, the Lagrangian function (6.21) is maximized when the city expands up to the point b at which the marginal benefit from both sides given by

$$2\{[Y - U^* + I + \lambda - \alpha \log n^o(b) - T(b)]n^o(b) - R_A\}$$

becomes zero. However, in so doing, we have not accounted for the fact that a marginal increase in b (respectively in $-b$) leads to an increase in the population by $n^o(b)$ at b (respectively at $-b$), which, in turn increases the travel costs for the other consumers by an amount equal to $n^o(b)T(b)$. Hence, the net benefit from expanding the city fringe is given by

$$2\{[Y - U^* + I + \lambda - \alpha \log n^o(b) - T(b)]n^o(b) - R_A\} - 2n^o(b)T(b).$$

Setting this expression equal to zero and using the optimality condition (6.22) evaluated at b, we obtain

$$\alpha n^o(b^o) - R_A = 0$$

so that

$$n^o(b^o) = \exp\left(\frac{\zeta^o - 2T(b^o)}{\alpha}\right) = \frac{R_A}{\alpha} \tag{6.24}$$

which is again identical to (6.9) except that $T(b^o)$ is now multiplied by 2.

Therefore, we may conclude that the equilibrium conditions become identical to the optimality conditions when $T(x)$ is replaced by $2T(x)$.

Using the same technique as in the equilibrium case, we can see that the optimum population density is given by

$$n^o(x) = \frac{\alpha}{2t}\frac{h^2 \exp(h|x|)}{(1 + \exp(h|x|))^2},$$

where h is a constant of integration that can be computed as k in the equilibrium case:

$$h^2 = \frac{t}{\alpha^2}(tN^2 + 8R_A).$$

It can then be readily verified that the optimum population density at the center is such that

$$n^o(0) = \frac{1}{\alpha}\left(\frac{tN^2}{8} + R_A\right).$$

Observe that the equilibrium and optimum solutions are identical when t is replaced by $2t$ in the former, that is, when consumers internalize the others' interaction costs. Because consumers have no incentives to do so, it appears that the optimum distribution is more concentrated than the equilibrium one. This result may come as a surprise because the conventional wisdom is that market cities are too crowded near the center. The reason for this surprising result is that consumers do not account for the locational externality they generate. Of course, we have not considered negative externalities (such as transport congestion or pollution) in the present model. Still, it is interesting to observe that *a preference for social life is sufficient to foster the emergence of an agglomeration of individuals* and that *the optimal agglomeration requires an even stronger concentration of people.* This is a fairly robust result, for we will encounter it below in very different models. In fact, it always holds provided that the individual benefit of interaction is additive across the whole population of agents (Fujita and Smith 1990).

6.2.2 The CBD as the Outcome of Interaction between Firms

In the same spirit, we now characterize the equilibrium distribution of office firms interacting together and using floor space. This topic has been addressed by Borukov and Hochman (1977) as well as by O'Hara (1977). We consider a continuum M of identical firms and a linear space for which the land density is 1, whereas the opportunity (presumably housing) land cost is $R_A > 0$. Each firm produces the same output Q_M, which is sold on a competitive market at a unit price. To do so, firms must interact with all other firms, thus bearing the corresponding transaction cost $T(x)$, and use one unit of floor space. This competition for proximity will lead firms to substitute capital for space under the form of office buildings. O'Hara (1977, 1196) put it this way:

> One firm's use of land near the center imposes higher travel costs on other firms which are precluded from using that land, and these costs are reflected in its rent. This provides an incentive to economize on these costs by substituting other inputs for land, especially by building taller buildings in the center than on the periphery of the CBD.

In other words, firms use floor space bought (or rented) on the office market at the prevailing office rent $R_o(x)$. Therefore, the profit of an office firm at location $x \in [-b, b]$ is defined as follows:

$$\pi(x) = Q_M - T(x) - R_o(x),$$

where the transaction cost borne by the same firm is

$$T(x) \equiv \int_{-b}^{b} t|x-y|m(y)dy \tag{6.25}$$

with $m(y)$ denoting the office firm density at y. As will be seen, the centrifugal force is expressed by the office rent, which is an increasing and convex function of the number of offices supplied at a given location.

Suppose that offices are supplied by the construction sector formed by a large number of competitive firms. If $S(x)$ denotes the amount of floor space provided by a developer per unit of land at location x, the developer's profit is supposed to be given by

$$\pi_c(x) = R_o(x)S(x) - [S(x)]^a - R(x),$$

where $[S(x)]^a$ stands for the construction cost of $S(x)$ units of floor space per unit of land and $R(x)$ is the land rent prevailing at location x; it is assumed that a is a constant greater than 1, and thus construction exhibits decreasing returns with respect to office height at the same place.

The equilibrium condition for construction firms implies that the density of floor space supplied at x is such that

$$S^*(x) = [R_o(x)/a]^{\frac{1}{a-1}}.$$

Thus, the supply of offices increases with the office rent. In turn, the market clearing condition for offices implies that $S^*(x) = m(x)$ for all locations x occupied by offices, thus showing that the equilibrium office rent must be such that

$$R_o(x) = a[m(x)]^{a-1}, \tag{6.26}$$

which is an increasing function of the number of firms established there since $a > 1$. As a result, at the free-entry equilibrium for the construction sector, we have

$$\pi_c = R_o(x)m(x) - [m(x)]^a - R(x) = 0,$$

and thus, using (6.26), we obtain

$$R(x) = (a-1)[m(x)]^a, \tag{6.27}$$

which means that the land rent prevailing at a particular location also increases with the number of firms set up there.

Because business firms are identical, in equilibrium each firm earns the same profit π^*, which is unknown but such that

$$R_o(x) = Q_M - T(x) - \pi^*$$

holds. Using (6.26), we obtain

$$a[m(x)]^{a-1} = Q_M - T(x) - \pi^*.$$

Differentiating this expression and using the relation $d^2T/dx^2 = 2tm(x)$, we obtain the following differential equation:

$$a(a-1)[m(x)]^{a-2}\frac{d^2m}{dx^2} + a(a-1)(a-2)[m(x)]^{a-3}\left(\frac{dm}{dx}\right)^2 + 2tm(x) = 0, \tag{6.28}$$

which is especially cumbersome because a need not be an integer. To simplify the analysis drastically, we assume that $a = 2$ so that (6.28) reduces to the following simple differential equation:

$$\frac{d^2m}{dx^2} + tm(x) = 0,$$

whose solution, that is, the equilibrium office firm density, is

$$m^*(x) = k \cos\left(t^{1/2}|x|\right),$$

where k is a positive constant of integration. In words, the equilibrium office firm density is therefore bell-shaped (as in Figure 6.1) with a maximum arising at $x = 0$, which is equal to k. Consequently, an information field is sufficient to explain the formation of a central business district even though the technology of the construction sector exhibits decreasing returns to scale at any given location. Furthermore, the density of offices, or the building height, decreases with the distance from the endogenous center inside the CBD, thus showing that the concentration of offices is the highest at the center.

From (6.27), where $a = 2$, the equilibrium land rent is

$$R^*(x) = k^2\left[\cos\left(t^{1/2}|x|\right)\right]^2,$$

which is also described by a bell-shaped curve as is the office rent obtained from (6.26):

$$R_o^*(x) = 2k \cos\left(t^{1/2}|x|\right).$$

At the fringe b of the business area, $R^*(b)$ must be equal to R_A, yielding

$$(R_A)^{1/2} = k \cos\left(t^{1/2}b\right). \tag{6.29}$$

Furthermore, the business firm population constraint implies that

$$M = \int_{-b}^{b} m(x)dx = 2k\int_0^b \cos\left(t^{1/2}x\right)dx,$$

which, after some manipulations, becomes

$$M = 2kt^{1/2} \sin\left(t^{1/2}b\right). \tag{6.30}$$

From (6.29) and (6.30), it then follows that

$$b^* = t^{-1/2} \arctan\left(\frac{M}{2}\sqrt{\frac{t}{R_A}}\right),$$

whereas k can be obtained by replacing b^* in (6.29):

$$k = \frac{(R_A)^{1/2}}{\cos\arctan\left(\frac{M}{2}\sqrt{\frac{t}{R_A}}\right)}.$$

Consequently, interaction among firms leads to the same kind of pattern within the CBD as interaction among consumers within the urban area. For exactly the same reason as the one discussed in the preceding section, the optimum density of firms must be more concentrated than the equilibrium density.[8]

6.3 THE CITY AS SPATIAL INTERDEPENDENCE BETWEEN FIRMS AND WORKERS

In Section 6.2, we have considered consumers' and firms' agglomeration separately. Here, we study how the interaction between both types of agents shapes the spatial structure of the entire city. Our discussion is based on a model in which the agglomeration force is generated, as in Section 6.2.2, through business externalities among firms, whereas social interaction among households is neglected for simplicity. Firms and households interact through perfectly competitive labor and land markets.

Specifically, the agglomeration force is due to the existence of communications among firms permitting the exchange of information. An important characteristic of information is that its transmission often requires direct communication between agents who typically incur distance-sensitive costs; hence, the benefits of information are larger when firms locate closer to each other. Therefore, all other things being equal, each firm has an incentive to establish itself close to other firms. On the other hand, the clustering of many firms into a single area increases the average commuting distance for workers which, in turn, gives rise to higher wages and land rent in the area surrounding the cluster. Such high wages and land rents tend to discourage further agglomeration of firms within the same area and act as a dispersion force. Consequently, the equilibrium distributions of firms and households/workers are determined as the balance between these two opposite forces.

6.3.1 The Model

Consider a one-dimensional space $X = (-\infty, \infty)$. The amount of land at each location $x \in X$ is equal to 1. There is a continuum N of homogeneous households/workers who are to reside in the city. There is also a continuum of *potential* firms that may operate in the city. Land is owned by absentee landlords, and firms by absentee shareholders. Workers are hired by active firms in the city. Both households and active firms use land. Land and labor markets are perfectly competitive at each and every location $x \in X$. The land rent prevailing at $x \in X$ is denoted by $R(x)$, and $W(x)$ stands for the wage at $x \in X$.

As usual, the utility of a household is given by $U(z, s)$, where s represents the land consumption and z the consumption of a composite good. For simplicity, it is assumed that the land consumption is fixed and equal to S_h. Furthermore, each household supplies one unit of labor, and the composite good is imported from outside the urban area at a constant price normalized to 1. Then, if a household chooses to reside at $x \in X$ and to work at $x_w \in X$, its budget constraint is given by

$$z + R(x)S_h + t|x - x_w| = W(x_w),$$

where t is the unit commuting cost.[9] Because the lot size is fixed, the objective of a household is to choose a residential location and a job site that maximize the consumption of the composite good given by

$$z(x, x_w) = W(x_w) - R(x)S_h - t|x - x_w|.$$

It is convenient to define a mapping J from X to X associating a (potential) job site $J(x) = x_w$ with a (potential) residential location x. This mapping describes the commuting pattern of workers and, for this reason, is called the *commuting function*. It follows that an individual residing at x must work at the location $J(x)$ that maximizes her net income:

$$W[J(x)] - t|x - J(x)| = \max_{y \in X}\{W(y) - t|x - y|\} \quad x \in X.$$

The associated *bid rent function* of a household at x is then defined as follows:

$$\Psi(x, u) = \{W[J(x)] - t|x - J(x)| - Z(u)\}/S_h, \tag{6.31}$$

where, as usual, $Z(u)$ is the solution to the equation $U(z, S_h) = u$. In this case, $\Psi(x, u)$ is the rent per unit of land that a household can bid at location x while working at $J(x)$ and enjoying the utility level u.

Firms produce the same good sold at a given price p and use the same technology. Specifically, each firm needs some fixed amount of land (S_f) and of labor (L_f) to undertake its production activity. However, the output level Q of a firm depends on the amount of information this firm obtains from the other firms in the city.

Firms are symmetric but different in the type of information they own. As a result, each firm wants to engage actively in communications with all other existing firms. The intensity of communications is measured by the level of contact activity (e.g., the number of face-to-face contacts), and each firm chooses its optimal level of contact activity with others. When $\varphi(x, y)$ is the level of contact activity chosen by a firm at $x \in X$ with a firm at $y \in X$, $V[\varphi(x, y)]$ represents the total contribution of this contact level to the firm's revenue pQ. This contribution, expressed by $V[\varphi(x, y)]$, is supposed to be the same across firms because of symmetry.

Communication from one firm to another is a pairwise activity that is time-consuming for both parties because of the need to organize, store, analyze, and communicate information. When a firm at x obtains information from a firm at y, the firm at x must bear a cost $c_1(x, y)$ per unit of contact, which is supposed to be a function of the location of the two firms. However, during this action, the firm at y also bears some cost c_2, which is typically independent of the firms' locations. For example, when a manager of a firm at x calls a manager of a firm at y, she imposes some cost on the manager contacted by consuming her time, and this cost does not depend on the interfirm distance. This means that each firm bears the additional cost generated by the communication activity taken by other firms.

Let $m(y)$ be the density of firms at location $y \in X$ and $\varphi(x, y)$ the level of contact activity chosen by a firm at $x \in X$ with each firm at y. Then, the revenue of a firm at x is given by

$$pQ(x) = \int_X \{V[\varphi(x, y)]\}m(y)dy,$$

and thus its profit is

$$\begin{aligned}\pi(x) &= pQ(x) - \int_X [c_1(x, y)\varphi(x, y) + c_2\varphi(y, x)]m(y)dy \\ &\quad - R(x)S_{\mathrm{f}} - W(x)L_{\mathrm{f}} \\ &= \int_X \{V[\varphi(x, y)] - c_1(x, y)\varphi(x, y) - c_2\varphi(y, x)\}m(y)dy \\ &\quad - R(x)S_{\mathrm{f}} - W(x)L_{\mathrm{f}}. \qquad (6.32)\end{aligned}$$

Each firm chooses its location x and its contact field $\varphi(x, y)$ so as to maximize its profit, taking the firm spatial distribution and contact field as given.[10]

The optimal contact level of a firm at x with any firm at y can be determined, independently of the whole distribution of firms, by choosing $\varphi(x, y)$ so as to maximize $V[\varphi(x, y)] - c_1(x, y)\varphi(x, y)$ in (6.32). If $c_1(x, y) = c_1(y, x)$, firms are symmetric in the process of communication so that the optimal level of contact between each pair of firms is the same for both of them: $\varphi^*(x, y) = \varphi^*(y, x)$.

We define the *local accessibility* between each location pair (x, y) by

$$a(x, y) \equiv V[\varphi^*(x, y)] - [c_1(x, y) + c_2]\varphi^*(x, y). \qquad (6.33)$$

Then the profit function (6.32) can be rewritten as follows:

$$\pi(x) = A(x) - R(x)S_{\mathrm{f}} - W(x)L_{\mathrm{f}}, \tag{6.34}$$

where

$$\begin{aligned} A(x) &\equiv \int_X a(x, y)m(y)dy \\ &= \int_X \{V[\varphi^*(x, y)] - [c_1(x, y) + c_2]\varphi^*(x, y)\}m(y)dy \end{aligned} \tag{6.35}$$

stands for the *aggregate accessibility* of each location $x \in X$.

Note that $a(x, y)$ could alternatively be interpreted as the information spillover experienced by a firm at x from a firm set up at y. In this case, $A(x)$ would represent the information field having the nature of a spatial externality. The amount of information received by a firm is in itself exogenous; however, it still depends on its location relative to the others.[11]

In association with (6.34), the *bid rent function* of a firm at x is defined as follows:

$$\Phi(x, \pi) = [A(x) - W(x)L_{\mathrm{f}} - \pi]/S_{\mathrm{f}}, \tag{6.36}$$

which represents the highest price a firm is willing to pay for a unit piece of land at $x \in X$ while earning a profit equal to π.

6.3.2 Equilibrium

As usual, the opportunity cost of land is given by R_{A}. The equilibrium configuration of the city is then determined through the interplay of the firms' and households' bid rent functions. More precisely, a *spatial equilibrium* is reached when all the firms achieve the same equilibrium profit π^*, all the households the same utility level given by u^*, and rents and wages clear the land and labor markets. The unknowns are the firm distribution $m(x)$, the household distribution $n(x)$, the land rent function $R(x)$, the wage function $W(x)$, the commuting function $J(x)$, and the equilibrium profit level π^* and utility level u^*.

For the N workers to be able to live in the city, all of them must get a job because they have to spend at least $R_{\mathrm{A}}S_{\mathrm{h}}$. Accordingly, there is full employment when

$$N = L_{\mathrm{f}}M, \tag{6.37}$$

and thus the equilibrium number of firms is

$$M^* = N/L_{\mathrm{f}}.$$

In our setting, because (1) the mass N of households is fixed, (2) each firm hires a fixed number of workers, and (3) there is no unemployment, either the equilibrium profit π^*, or the equilibrium utility level u^*, is indeterminate.

Here, we assume free entry and exit of firms so that $\pi^* = 0$ is an additional equilibrium condition.

From the bid rent functions (6.31) and (6.36), the equilibrium conditions can be described as follows:

1. *land market equilibrium*: at each $x \in X$,

$$R(x) = \max\{\Psi(x, u^*), \Phi(x, 0), R_A\} \tag{6.38}$$

$$\Psi(x, u^*) = R(x) \quad \text{if } n(x) > 0 \tag{6.39}$$

$$\Phi(x, 0) = R(x) \quad \text{if } m(x) > 0 \tag{6.40}$$

$$S_h n(x) + S_f m(x) = 1 \quad \text{if } R(x) > R_A, \tag{6.41}$$

2. *commuting equilibrium*: at each $x \in X$,

$$W[J(x)] - t|J(x) - x| = \max_{y \in X}[W(y) - t|y - x|] \tag{6.42}$$

3. *labor market equilibrium*: at each $x \in X$,

$$\int_I n(x)dx = \int_{J(I)} L_f m(x)dx \quad \text{for every interval } I \text{ of } X \tag{6.43}$$

4. *firms' and households' population constraints*:

$$\int_X m(x)dx = M^* = N/L_f \tag{6.44}$$

$$\int_X n(x)dx = N. \tag{6.45}$$

Conditions (6.38) through (6.41) together mean that each location is occupied by agents with the highest bid rent. Condition (6.42) says that, for each potential residential location x, the commuting destination $J(x)$ maximizes the net wage. Condition (6.43) ensures the equality of labor supply and demand under the commuting function J. The meaning of the population conditions is obvious.

It is worth noting one general property of any spatial equilibrium. Given a commuting function J, if there exist $x \in X$ and $x' \in X$ such that

$$(J(x) - x)(J(x') - x') < 0 \quad \text{and} \quad (x - x')(J(x) - J(x')) < 0, \tag{6.46}$$

we say that *cross-commuting* occurs. The first inequality means that a resident at x and a resident at x' commute in the opposite directions, whereas the second one implies that their commuting paths have an overlapping section. Because (6.42) requires that each resident chooses a job site that maximizes her net income, the following result is intuitively obvious: in any spatial equilibrium configuration, cross-commuting does not occur. Indeed, if two groups of workers cross-commute, any household belonging to any of these groups would strictly increase its net income by choosing a job site in the area in which the other group works.

It turns out that the properties of the equilibrium urban configuration crucially depend on the shape of the local accessibility function $a(x, y)$. By specifying the analytical form of the benefit function $V(\varphi)$ and of the cost function $c_1(x, y)$, we can obtain a different expression for $a(x, y)$. In what follows, we focus on two special but meaningful cases:

$$a(x, y) = \beta - \tau|x - y| \tag{6.47}$$

and

$$a(x, y) = \beta \exp(-\tau|x - y|), \tag{6.48}$$

where τ and β are two positive constants, τ measuring the intensity of the distance-decay effect. The former equation corresponds to a *linear accessibility*, whereas the latter is a *spatially discounted accessibility*. Both expressions have been used extensively in spatial models of interaction.

For example, expression (6.48) can be justified on the following grounds. Let us assume that the benefit function is given by the *entropy-type function*

$$V(\varphi) = \begin{cases} -\varphi \log \varphi & \text{for } \varphi < 1/e \\ 1/e & \text{for } \varphi \geq 1/e \end{cases}$$

representing the firms' propensity to collect heterogeneous information in the context of (6.32). Then, for each location pair (x, y), by choosing $\varphi(x, y)$ so as to maximize $\{V[\varphi(x, y)] - c_1(x, y)\varphi(x, y)\}$, we obtain the optimal level of contact

$$\varphi^*(x, y) = \exp[-1 - c_1(x, y)].$$

Substituting this expression into identity (6.33) yields

$$a(x, y) = (1 - c_2)\exp[-1 - c_1(x, y)]$$

or, setting $\beta \equiv (1 - c_2)/e$, which is assumed to be positive, yields

$$a(x, y) = \beta \exp[-c_1(x, y)].$$

If $c_1(x, y) = \tau|x - y|$, we obtain (6.48).[12]

In the case of linear accessibility, Ogawa and Fujita (1980) and Imai (1982) have shown independently that a unique equilibrium configuration exists for each parameter constellation. Furthermore, in this case, only three possible equilibrium configurations exist that are all essentially monocentric. In contrast, in the case of spatially discounted accessibility, Fujita and Ogawa (1982) have established multiplicity of equilibria. Furthermore, some equilibrium configurations display several centers. In the following two sections, we examine each case in turn.

6.4 THE MONOCENTRIC CITY

When the local accessibility is linear (6.47), the aggregate accessibility of each location $x \in X$ is given by:

$$A(x) = \int_X [\beta - \tau|x - y|]m(y)dy. \tag{6.49}$$

Because the support of the density m is not necessarily connected, its median, denoted $med[m]$, may be an interval. Hence,

$$\frac{dA(x)}{dx} = -\tau\left\{\int_{-\infty}^{x} m(y)dy - \int_{x}^{\infty} m(y)dy\right\}, \tag{6.50}$$

which is positive, zero, or negative, respectively, as $x < \inf med[m]$, $x \in med[m]$, or $x > \sup med[m]$. Furthermore, the support of the density m, denoted $m_+ \equiv \{x \,|\, m(x) > 0\}$, is the *business area*, and we have

$$\frac{d^2A(x)}{dx^2} = -2\tau m(x)\begin{cases} < 0 & x \in m_+ \\ = 0 & x \notin m_+ \end{cases}, \tag{6.51}$$

implying that $A(x)$ is globally concave, that is, strictly concave over the business area and linear elsewhere. That $A(x)$ is globally concave creates a strong agglomeration force, thus suggesting that firms do not want to be too far apart from one another.

In what follows, we determine each potential equilibrium configuration in turn and then examine under which conditions they are equilibria using the basic principles of land competition described in Chapter 3. As will be seen, the bid rent functions are continuous with respect to x so that all the intervals presented may be considered as closed.

We first represent in Figure 6.2 the monocentric city case. The upper diagram depicts the corresponding land use pattern together with the associated wage curve $W(x)$. There is a central business district, defined by the interval $[-b_1, b_1]$, surrounded by two residential sections of equal size, $[-b_2, -b_1]$ and $[b_1, b_2]$. Because there is no cross-commuting, the left (right) side half of the business area contains the job sites associated with the locations belonging to the left (right) side residential section. For all workers to have the same net wage, the wage curve must be linear on each side of the city center and must have a slope equal to the unit commuting cost t. This is an equilibrium if and only if the bid rent curves associated with the monocentric pattern are as those depicted in the lower diagram of Figure 6.2, that is, the households' bid rent curve dominates the firms' bid rent curve in the business area, whereas the converse holds in each section of the residential area.

In the monocentric configuration, the business area is completely separated from the residential area so that each worker must commute to her job site. At the other extreme, the completely integrated configuration is depicted in

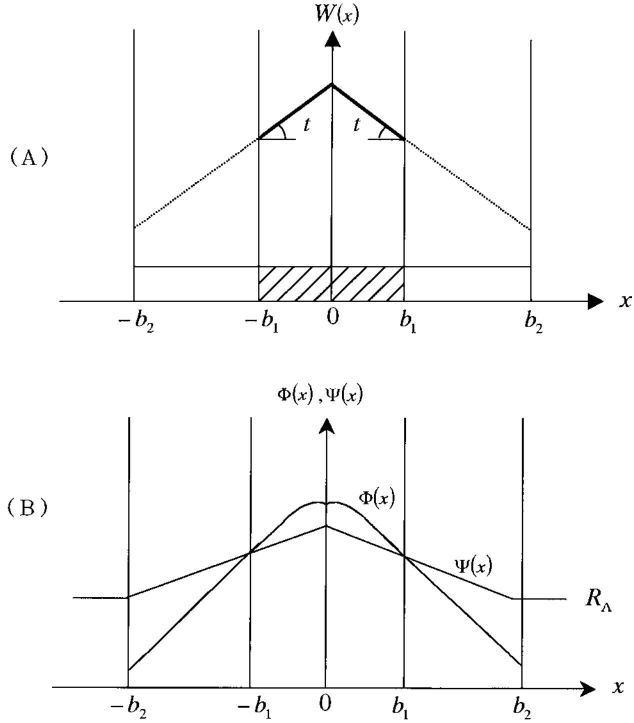


Figure 6.2: The monocentric configuration.

Figure 6.3, where both firms and households are uniformly distributed inside the urban area $[-b_2, b_2]$. In this case, the job site is identical to the residential location for all workers so that there is no commuting. For this to be a spatial equilibrium, the corresponding households' and firms' bid rents must be equal at each location.

The third case corresponds to the incompletely integrated configuration depicted in Figure 6.4 and is a mixture of the first two cases. As shown in the upper diagram, firms and households are uniformly mixed within the *integrated district*, $[-b_0, b_0]$, where no commuting arises. This area is surrounded by two business sections, $[-b_1, -b_0]$ and $[b_0, b_1]$, each of which is adjacent to a residential section, $[-b_2, -b_1]$ and $[b_1, b_2]$. The bid rent curves supporting such a configuration must resemble those given in the lower diagram of Figure 6.4.

The reader has probably noticed that the first two configurations are special cases of the third one. Indeed, if we set $b_0 = 0$ in Figure 6.4, we obtain the monocentric pattern; similarly, if we set $b_0 = b_2$ we get the completely integrated configuration. In other words, the incompletely integrated configuration

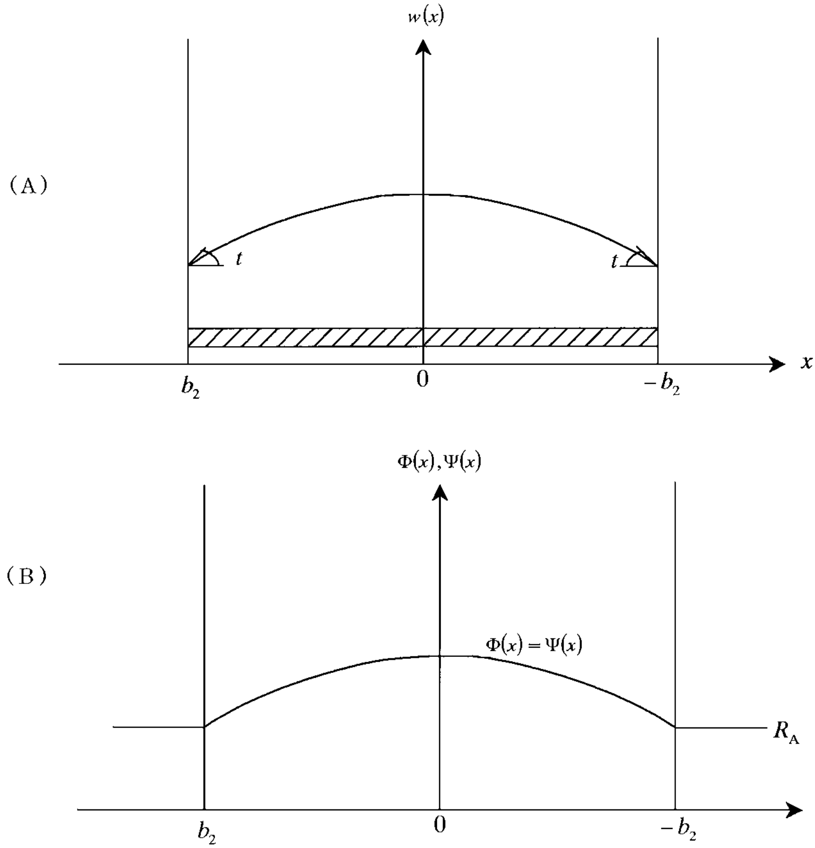


Figure 6.3: The completely integrated configuration.

corresponds to the generic pattern under linear accessibility. Consequently, in what follows, we first determine the set of parameters for the incompletely integrated configuration to be an equilibrium. The other two configurations may then be derived as special cases. In so doing, we will see that the whole domain of admissible values for the parameters is fully covered.

Consider the configuration represented in Figure 6.4. Because the configuration is symmetric about $x = 0$, we may restrict ourselves to the nonnegative values of x. In the integrated area, $[0, b_0]$, all individuals work at their residential place, thus implying that $J(x) = x$ for all $x \in [0, b_0]$. The land constraint condition (6.41) and the labor market clearing condition (6.43) then imply

$$m(x) = 1/(S_f + S_h L_f) \quad \text{and} \quad n(x) = L_f/(S_f + S_h L_f) \quad x \in [0, b_0]. \tag{6.52}$$

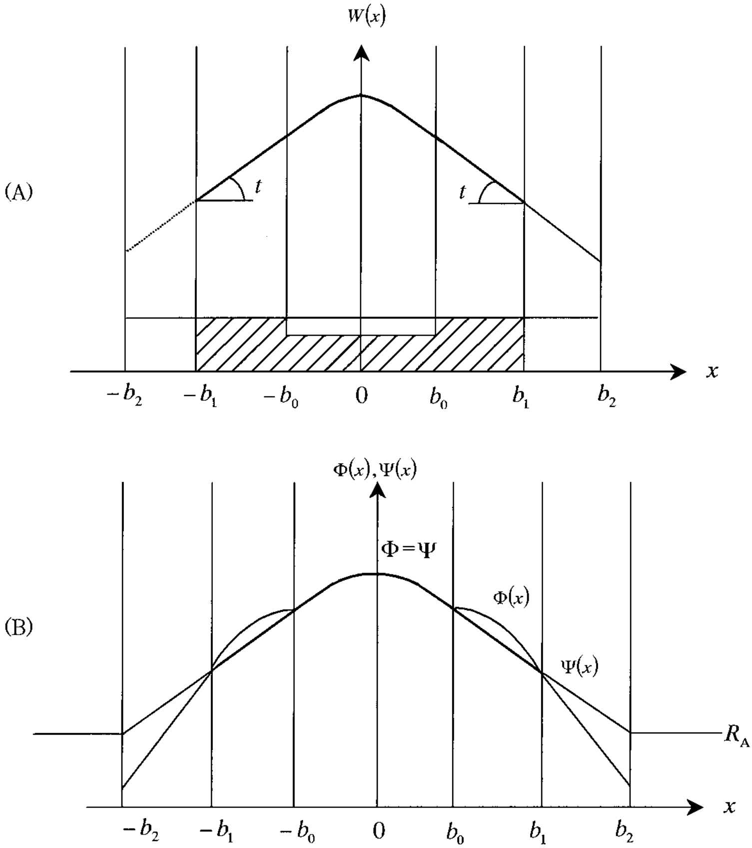


Figure 6.4: The incompletely integrated configuration.

In each business section, it is necessary that

$$m(x) = 1/S_{\mathrm{f}} \quad \text{and} \quad n(x) = 0 \quad x \in [b_0, b_1]. \tag{6.53}$$

Likewise, in each residential section, we must have

$$n(x) = 1/S_{\mathrm{h}} \quad \text{and} \quad m(x) = 0 \quad x \in [b_1, b_2]. \tag{6.54}$$

In addition, the population constraint (6.44) implies that

$$\frac{M^*}{2} = \int_0^{b_0} m(x)dx + \int_{b_0}^{b_1} m(x)dx = \frac{b_0}{S_{\mathrm{f}} + S_{\mathrm{h}}L_{\mathrm{f}}} + \frac{b_1 - b_0}{S_{\mathrm{f}}},$$

which yields the equilibrium boundary location as a function of b_0:

$$b_1^*(b_0) \equiv b_1^* = \frac{S_{\mathrm{h}}L_{\mathrm{f}}}{S_{\mathrm{f}} + S_{\mathrm{h}}L_{\mathrm{f}}} b_0 + \frac{S_{\mathrm{f}}M^*}{2}. \tag{6.55}$$

Because there is no vacant land between $-b_2^*$ and b_2^*, it can readily be verified that $MS_\mathrm{f} + NS_\mathrm{h} = 2b_2^*$ implies that the city fringe is as follows:

$$b_2^* = (S_\mathrm{f} + S_\mathrm{h} L_\mathrm{f})M/2. \tag{6.56}$$

Since no cross-commuting occurs in the integrated district and all the individuals residing in the section $[b_1^*, b_2^*]$ work in the business section $[b_0, b_1^*]$, we can assume without loss of generality that the commuting function $J(x)$ is as follows:

$$J(x) = \begin{cases} x & x \in [0, b_0] \\ \frac{b_1^*(x - b_1^*) + b_0(b_2^* - x)}{b_2^* - b_1^*} & x \in [b_0, b_2^*] \end{cases}, \tag{6.57}$$

where $J(b_1^*) = b_0$ and $J(b_2^*) = b_1^*$. This means that an individual living in the integrated district $[0, b_0]$ is assigned to her residential location, whereas one living in the residential section $[b_1^*, b_2^*]$ is assigned to a location belonging to the business section $[b_0, b_1^*]$ with x and $J(x)$ varying in the same direction. To support this commuting pattern, condition (6.42) requires[13]

$$W^*(x) = W^*(b_0) - t(x - b_0) \quad x \in [b_0, b_2^*]. \tag{6.58}$$

Next, substituting (6.52)–(6.54) into (6.49), we obtain the following description of the aggregate accessibility:

$$\begin{aligned} A^*(x) &= \beta M^* - \left\{ \frac{\tau}{S_\mathrm{f}}\left(b_1^2 - b_0^2\right) + \frac{\tau}{S_\mathrm{f} + S_\mathrm{h} L_\mathrm{f}}\left(b_0^2 + x^2\right) \right\} \quad x \in [0, b_0] \\ &= \beta M^* - \left\{ \frac{\tau}{S_\mathrm{f}}\left(b_1^2 - 2b_0 x + x^2\right) + \frac{2\tau}{S_\mathrm{f} + S_\mathrm{h} L_\mathrm{f}} b_0 x \right\} \quad x \in [b_0, b_1^*] \\ &= \beta M^* - \left\{ \frac{2\tau}{S_\mathrm{f}}(b_1 - b_0 x) + \frac{2\tau}{S_\mathrm{f} + S_\mathrm{h} L_\mathrm{f}} b_0 x \right\} \quad x \in [b_1^*, b_2^*]. \end{aligned} \tag{6.59}$$

With $\pi^* = 0$ in (6.36), the equilibrium bid rent function of a firm is

$$\Phi^*(x) \equiv \Phi(x, 0) = [A^*(x) - W^*(x) L_\mathrm{f}]/S_\mathrm{f}. \tag{6.60}$$

Similarly, setting $u = u^*$ in (6.31) and using the first part of (6.57) as well as (6.58), we obtain the equilibrium bid rent function of a household:

$$\begin{aligned} \Psi^*(x) \equiv \Psi(x, u^*) &= [W^*(x) - Z(u^*)]/S_\mathrm{h} \quad x \in [0, b_0] \\ &= [W^*(b_0) - t(x - b_0) - Z(u^*)]/S_\mathrm{h} \quad x \in [b_0, b_2^*]. \end{aligned} \tag{6.61}$$

As shown by the lower diagram of Figure 6.4, the incompletely integrated configuration is a spatial equilibrium if and only if (1) the commuting condition (6.42) holds and (2) the equilibrium bid rent functions satisfy the following conditions:[14]

$$\Phi^*(x) = \Psi^*(x) \quad x \in [0, b_0] \tag{6.62}$$

$$\Phi^*(x) \geq \Psi^*(x) \quad x \in [b_0, b_1^*] \tag{6.63}$$

$$\Phi^*(x) \leq \Psi^*(x) \quad x \in [b_1^*, b_2^*] \tag{6.64}$$

$$\Psi^*(b_2^*) = R_A \tag{6.65}$$

with the commuting function given by (6.57).

The remainder of the argument involves five more steps.

1. Using (6.56), we obtain from the second part of (6.61) and from (6.65):

$$W^*(b_0) = R_A S_h + t[(S_f + S_h L_f)M - 2b_0]/2 + Z(u^*). \tag{6.66}$$

2. Using (6.62) together with (6.60) and (6.61) yields the equilibrium wage inside the integrated area:

$$W^*(x) = \frac{A^*(x)S_h + Z(u^*)S_f}{S_f + S_h L_f} \quad x \in [0, b_0]. \tag{6.67}$$

   We evaluate (6.67) by substituting $A^*(x)$ as defined by the first equality in (6.59) into (6.67) and using (6.55). The value of $W^*(x)$ at b_0 may then be equalized to (6.66) to determine u^* uniquely as a function of b_0, denoted $u^*(b_0)$. From (6.58), (6.66), and (6.67), the equilibrium wage $W^*(x)$ at each location $x \geq 0$ may be uniquely determined as a function of b_0.

3. To determine b_0, we observe that (6.62) implies

$$\Phi^*(b_0) = \Psi^*(b_0), \tag{6.68}$$

   whereas (6.63) and (6.64) lead to

$$\Phi^*(b_1^*) = \Psi^*(b_1^*). \tag{6.69}$$

   These two conditions then yield

$$\frac{\tau}{S_f}(b_1^* - b_0) + \frac{2\tau b_0}{S_f + S_h L_f} = \frac{(S_f + S_h L_f)t}{S_h}.$$

   Substituting b_1^* given by (6.55) into this expression allows for the determination of the equilibrium value of b_0, which is given by:

$$b_0^* = \frac{t(S_f + S_h L_f)^2}{\tau S_h} - \frac{(S_f + S_h L_f)M^*}{2}. \tag{6.70}$$

   This means that b_0^* is a linear function of t. Because b_0^* must belong to the interval $[0, b_2^*]$, where b_2^* is given by (6.56), the following inequalities must hold:

$$\frac{S_h M^*}{2(S_f + S_h L_f)} \leq \frac{t}{\tau} \leq \frac{S_h M^*}{S_f + S_h L_f}. \tag{6.71}$$

4. It is easy to check that $\Psi^*(x)$ is linear for $x \in [b_0^*, b_2^*]$, that $\Phi^*(x)$ is strictly concave for $x \in [b_0^*, b_1^*]$ and linear for $x \in [b_1^*, b_2^*]$ while the left- and right-hand side derivatives of $\Phi^*(x)$ are equal at b_1^*. These properties, together with (6.68) and (6.69), imply immediately that (6.63) and (6.64) hold.
5. It remains to check that the commuting function defined by (6.57) is sustainable. Because $W^*(x)$ is strictly concave on $[0, b_1^*]$ and linear on $[b_1^*, b_2^*]$ with slope $-t$, the commuting pattern is part of the equilibrium if and only if the absolute value of the left-hand side derivative of $W^*(x)$ at b_0^* does not exceed t. From the first equality of (6.59) and (6.67), this means

$$t \geq \frac{2\tau S_h b_0^*}{(S_f + S_h L_f)^2} \tag{6.72}$$

or, using (6.70),

$$t \geq 2t - \frac{\tau S_h M^*}{S_f + S_h L_f},$$

which amounts to

$$\frac{t}{\tau} \leq \frac{S_h M^*}{S_f + S_h L_f}.$$

This condition is satisfied whenever condition (6.71) holds.

Consequently, using (6.37), we have shown the following result:

Proposition 6.2 *The incompletely integrated configuration is a spatial equilibrium if and only if*

$$\frac{S_h N}{2(S_f + S_h L_f)L_f} \leq \frac{t}{\tau} \leq \frac{S_h N}{(S_f + S_h L_f)L_f}.$$

As noted above, the incompletely integrated configuration degenerates into the monocentric configuration represented in Figure 6.2 once $b_0^* = 0$. By setting $b_0 = 0$ in all the analyses above and by replacing (6.68) by

$$\Phi^*(b_0) \geq \Psi^*(b_0)$$

we obtain:[15]

Proposition 6.3 *The monocentric configuration is a spatial equilibrium if and only if*

$$\frac{t}{\tau} \leq \frac{S_h N}{2(S_f + S_h L_f)L_f}.$$

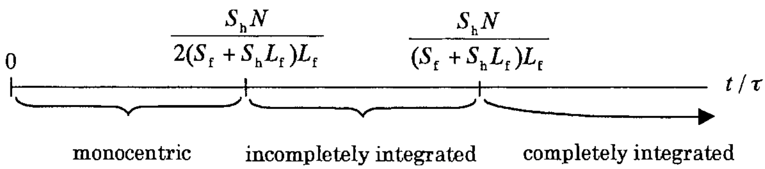


Figure 6.5: The parameter ranges for the three equilibrium configurations under linear accessibility.

In the same manner, by setting $b_0 = b_2$ in the preceding developments and by replacing b_0^* in (6.72) by b_2^* given in (6.56), we get[16]

Proposition 6.4 *The completely integrated configuration is a spatial equilibrium if and only if*

$$\frac{t}{\tau} \geq \frac{S_h N}{(S_f + S_h L_f) L_f}.$$

All the results above are summarized in Figure 6.5 in which the whole domain of possible values for t/τ is covered. Among other things, we see that the monocentric (respectively completely integrated) configuration is an equilibrium when t, the unit commuting cost, is relatively small (respectively large) in comparison with τ, the distance-decay parameter in communication, and N, the city size. Hence, as commuting costs fall while the intensity of communication between firms rises (two fairly general trends observed since the Industrial Revolution), one moves from backyard capitalism to a monocentric city with complete specialization of land. This means that the monocentric configuration is likely to emerge when commuting costs are low, when the spatial distance-decay effect is strong, or both. If the second result is fairly intuitive, the first is probably less apparent but will appear on several occasions in subsequent chapters, for low transportation costs seem foster agglomeration. On the other hand, high transportation costs lead to the completely mixed configuration, that is, a pattern with no land specialization and no commuting. All these results confirm and extend those we have found in Section 3.2.2.[17]

These three propositions may also be interpreted in terms of local labor markets. In the case of a monocentric city, a single labor market is established in the central business district. When the equilibrium configuration is completely integrated, labor is traded in each location within the city. The pattern of labor markets is more involved in the case of an incompletely integrated configuration: labor is locally traded in the integrated district, which is surrounded by two local labor markets attracting workers from the city outskirts.

Showing the uniqueness of these configurations involves fairly elaborate arguments (Ogawa and Fujita 1980). For simplicity, we will restrict ourselves

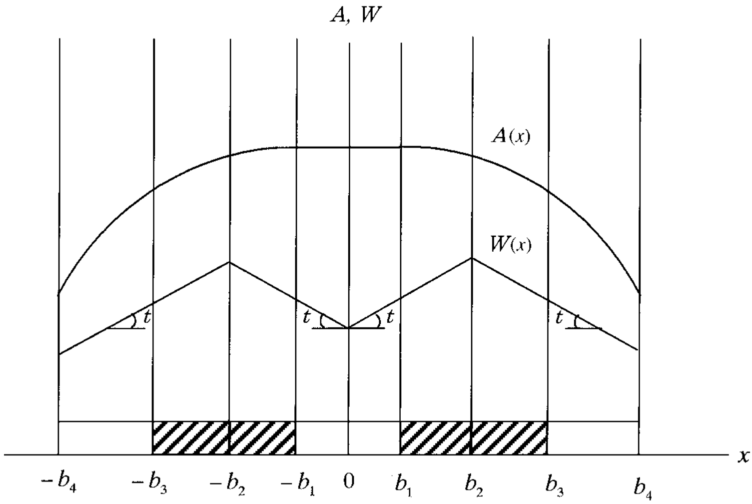


Figure 6.6: The impossibility of a duocentric configuration under linear accessibility.

to the case of a duocentric configuration and will give an intuitive argument showing why it cannot be an equilibrium under linear accessibility.

Consider the symmetric duocentric configuration depicted in Figure 6.6 in which each of the business areas $[b_1, b_3]$ and $[-b_3, -b_1]$ is surrounded by two residential sections from which the required labor is supplied. To support such a commuting pattern, the wage function must take the following form:

$$W^*(x) = W^*(b_2) - t|b_2 - x| \quad x \in [0, b_4],$$

which shows a peak at some location b_2 between b_1 and b_3. By (6.31) and (6.36), this means that the equilibrium bid rent functions are now given by

$$\Phi(x, 0) = \{A(x) - [W^*(b_2) - t|b_2 - x|]L_\mathrm{f}\}/S_\mathrm{f} \quad x \in [0, b_4]$$
$$\Psi(x, u^*) = [W^*(b_2) - t|b_2 - x| - Z(u^*)]/S_\mathrm{h} \quad x \in [0, b_4].$$

Furthermore, since b_1 is by definition the boundary between a business district and a residential section, it must be that

$$\Phi(b_1, 0) = \Psi(b_1, u^*).$$

Because $|b_2 - x| = b_2 - x$ for $x \in [0, b_1]$, the three expressions above imply that

$$\begin{aligned}\Phi(x, 0) - \Psi(x, u^*) &= [\Phi(x, 0) - \Phi(b_1, 0)] - [\Psi(x, u^*) - \Psi(b_1, u^*)]\\ &= [A^*(x) - A^*(b_1)]/S_\mathrm{f}\\ &\quad + t(b_1 - x)(S_\mathrm{f} + S_\mathrm{h}L_\mathrm{f})/(S_\mathrm{f} + S_\mathrm{h}).\end{aligned}$$

However, in as much as $dA(x)/dx = 0$ by (6.50) for $x \in [0, b_1]$, $A^*(x) = A^*(b_1)$ for $x \in [0, b_1]$ and thus we have

$$\Phi(x, 0) > \Psi(x, u^*) \quad x \in [0, b_1],$$

a result contradicting the assumption that $[0, b_1]$ is part of a residential section.

In other words, when the local accessibility is linear, the aggregate accessibility is *flat* on the central interval $[-b_1, b_1]$. Consequently, because the wage function decreases from $x = b_1$ to $x = 0$, firms can afford to make higher bids inside this residential area than they do at b_1, whereas households can only make lower bids than they do at b_1. That the two bids are equal at b_1 implies firms would choose to be in $[-b_1, b_1]$.

Figure 6.6 also shows that, for a duocentric configuration to be an equilibrium, the aggregate accessibility function $A(x)$ must exhibit a peak inside each business district as the wage function $W(x)$ does. This is possible provided that the local accessibility function $a(x, y)$ decreases sufficiently fast with the distance between x and y, which may occur with the spatially discounted accessibility (6.48).

Finally, let us briefly discuss the optimal configuration under linear accessibility.[18] As in the models presented in Section 6.2, the competitive equilibrium is inefficient because a firm, when choosing a location, considers only its own communication costs and disregards the change in the same cost incurred by the other firms.[19] For the optimal configuration to emerge as an equilibrium, each firm should internalize the communication costs of each other firm, that is, τ should be replaced by 2τ in all the formal developments presented above. Thus, Figure 6.5 can be replaced by Figure 6.7, which specifies the parameter ranges corresponding to the different optimal configurations.

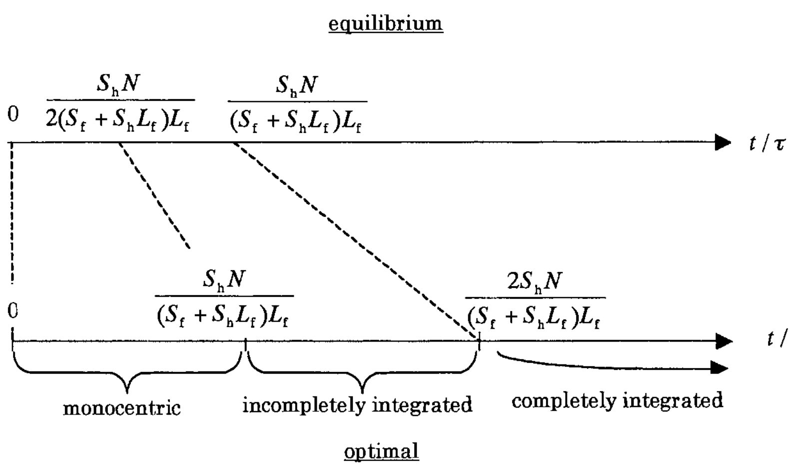


Figure 6.7: Parameter ranges for the three socially optimal configurations under linear accessibility.

As a consequence, the domain of t/τ for which the monocentric configuration is socially optimal is twice as large as the domain for which it is an equilibrium. Again, as in the preceding section, firms tend to be less agglomerated in equilibrium than they should be at the optimum.

6.5 THE POLYCENTRIC CITY

The set of market outcomes under the spatially discounted accessibility is much richer, but also more complex, than under linear accessibility. Indeed, differentiating $A(x)$ twice, in which $a(x, y)$ is given by (6.48), leads to

$$\frac{d^2 A(x)}{dx^2} = -\beta\tau[2m(x) - \tau A(x)].$$

Given that $A(x) < 2/\tau S_{\mathrm{f}}$ at any x (see Part C of the chapter appendix), it then follows that $A(x)$ is strictly convex on any residential section, because $m(x) = 0$, and strictly concave on any business section given that $m(x) = 1/S_{\mathrm{f}}$. This implies that $A(x)$ may display several peaks and, in turn, opens the door to the possible emergence of several employment centers. Furthermore, the impact of the parameter τ on the agglomeration force is not monotone. Indeed, it is obvious that $A(x)$ is flat and equal to M (the total number of firms) when $\tau = 0$, and $A(x)$ is again flat but equal to 0 when $\tau \to \infty$. The function $A(x)$ is depicted in Figure 6.8 for different values of τ in the case of a single business district. For example, differences in accessibility at the center and at the fringe of the business district are greatest for intermediate values of τ. The profit function (6.34) shows that such differences are critical in the choice of a location by a firm. Accordingly, we may expect to observe new results for the intermediate

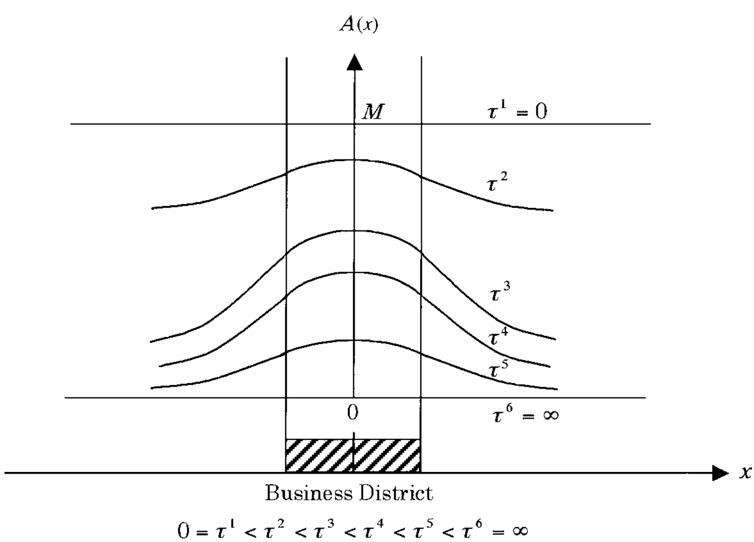


Figure 6.8: The impact of τ on the accessibility function $A(x)$.

values of τ. However, the counterpart of the richness of the results is a much higher level of analytical complexity. As a result, we will be content to appeal to numerical analysis when necessary.

When τ is very small, it should be obvious that the spatially discounted accessibility can be well approximated by the linear one, and thus the three possible configurations described in the preceding section still prevail in such cases.[20] As we just pointed out, the more involved situations occur when τ takes intermediate values. In such cases, the following may occur: (1) there exists configurations exhibiting several centers; (2) there is often multiplicity of equilibria; and (3) the transition from an equilibrium to another may be catastrophic.

Consider first the emergence of a polycentric city. In addition to the three configurations provided in Figures 6.2–6.4, we describe in Figures 6.9–6.11 the most typical examples of what may be observed. In Figure 6.9, a duocentric city with two business districts of equal sizes is shown. This pattern may also be interpreted as two adjoining cities creating external economies for each other and, therefore, enjoying agglomeration economies within a system of cities.

In Figure 6.10, the equilibrium city has one primary center and two secondary centers of identical size. The arrows indicate the direction of the workers' commuting flows. Observe that some workers cross a secondary center because they work in the primary center. This is because firms located in a secondary center cannot be too far from those in the primary center to enjoy the external effects generated by this center. In terms of labor markets, this means that the primary center attracts workers from all the residential areas whereas each secondary center pulls people from its periphery only.

Last, in Figure 6.11, there are three approximately identical centers surrounded by two residential sections each, which is a pattern akin to three connected small cities. The two internal residential sections are occupied by people who work in the middle business section or in one of the two peripheral business sections. Each center attracts workers from its two neighboring residential sections only.

It must be stressed that these configurations do not exhaust the set of equilibria. Because of the nonlinearities arising in the model, it is hard to provide a full characterization of this set.

We now come to the problem of multiplicity of equilibria. In Figure 6.12, drawn from Fujita and Ogawa (1982), the regions in the parameter space (t, τ) corresponding to the various equilibria described earlier are depicted for the parameter values $(N, S_f, S_h, L_f) = (1000, 1, 0.1, 10)$ so that firms (households) use 100 units of land. Below the locus denoted as C_I lie the monocentric configurations; above C_C is the domain of completely integrated configurations; the domain of incompletely integrated configurations can be found between C_I and C_C; the domain delineated by the broken line C_2 is that of duocentric configurations, whereas the domain delineated by the solid line C_{3A} (C_{3B}) corresponds to the cities with one center and two secondary centers of identical sizes (three approximately identical centers).

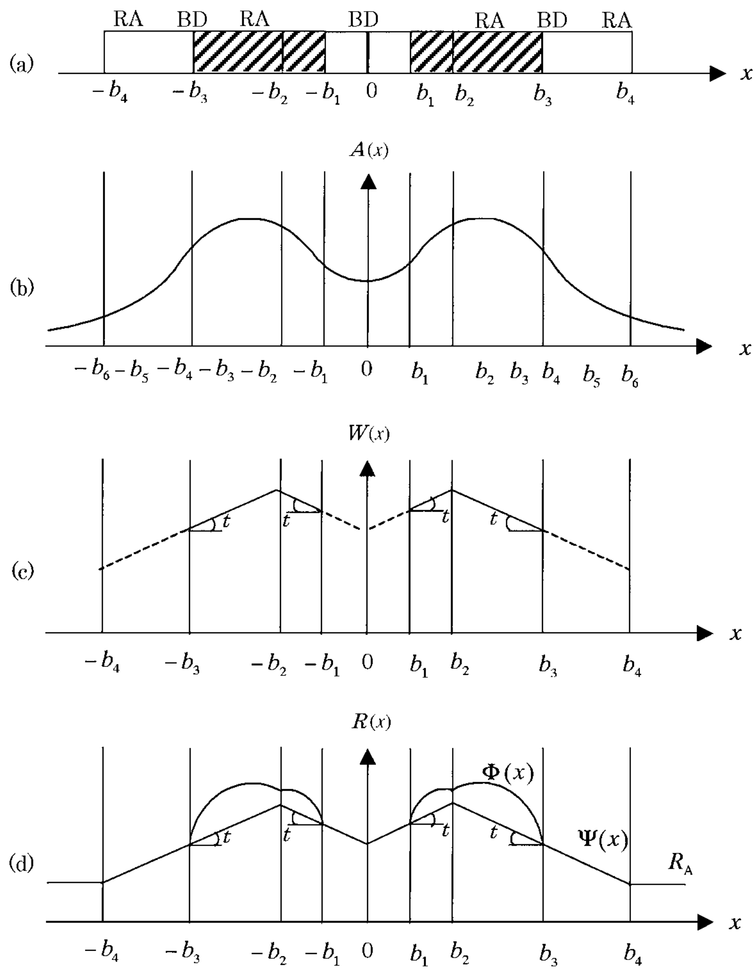


Figure 6.9: The duocentric configuration under the spatially discounted accessibility.

It is easily seen that some of these regions overlap, and thus, for the same parameter configuration, several equilibrium patterns may exist. For example, in the shaded area, four equilibria exist: the monocentric, incompletely integrated, duocentric, and tricentric (with two subcenters) configurations. Note, however, that we do not know anything about the stability of these equilibria. Performing such an analysis when the unknowns are continuous curves is a hard task left for future research.

Finally, it should be noted that small changes in some parameters may lead to dramatic modifications in the prevailing equilibrium pattern. For example, if $t = 0.007$ in Figure 6.12, the following path may well arise. When τ is very small, we have an integrated city because commuting costs are dominant; above

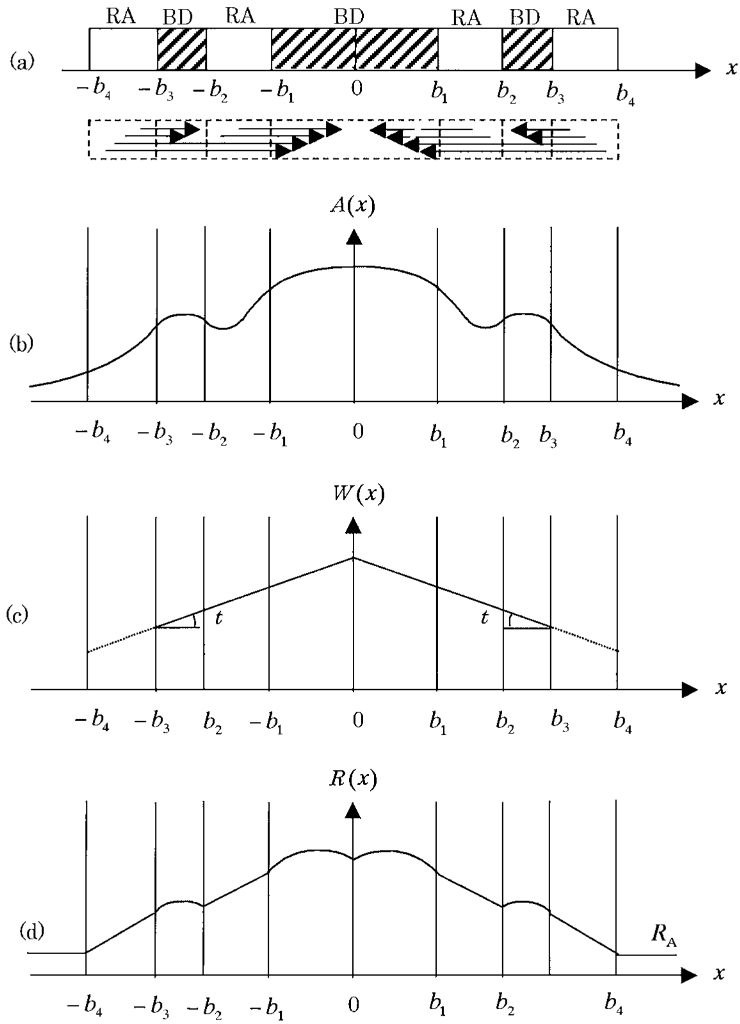


Figure 6.10: The tricentric configuration with one primary center and two secondary centers.

some threshold, the city is incompletely integrated because the accessibility of each firm to the others now matters; when τ keeps rising, we reach the domain in which the city is monocentric. So far, the transition is smooth and similar to that observed under linear accessibility. The picture changes drastically when τ takes some intermediate values larger than a, as shown in Figure 6.10. The city becomes duocentric. At this point, the size of the central residential section is significantly large ($\cong$50 units of land). Similarly, the aggregate differential rent drops down by a fairly large amount because a much lower degree of centrality now prevails within the city. The transition at point a is therefore catastrophic.[21]

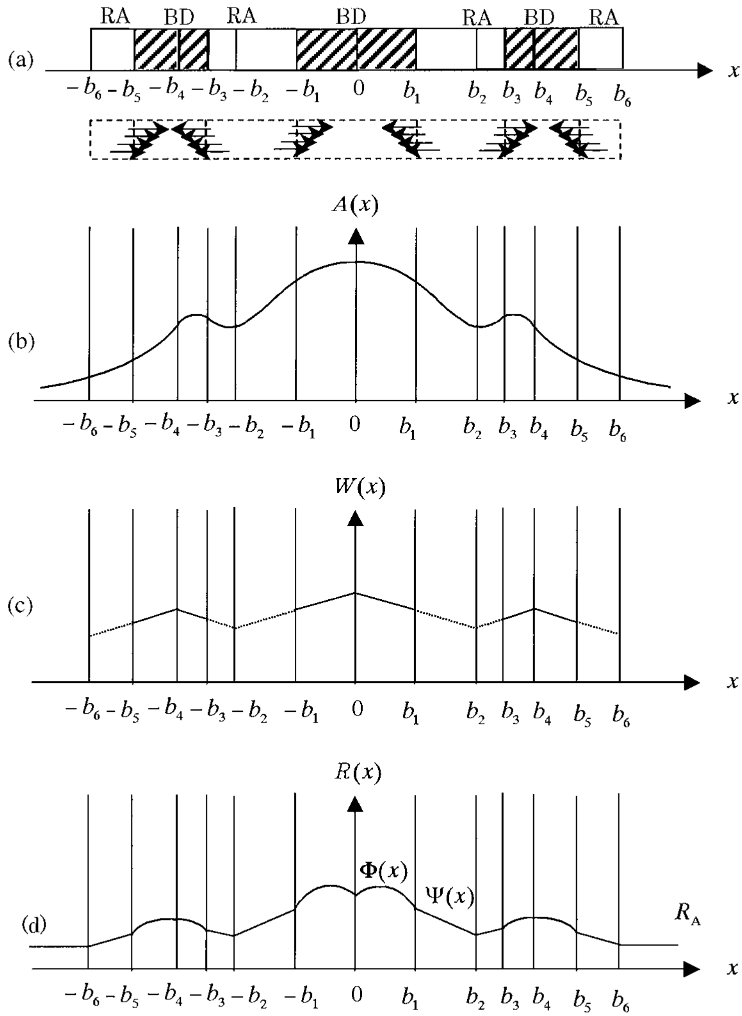


Figure 6.11: The tricentric configuration with three approximately identical centers.

Interestingly, the equilibrium is also path-dependent because it shows inertia. For example, decreasing τ from the right of a yields a duocentric pattern up to point b, where the size of the central residential section is approximately equal to 30 units of land. At this point, the city exhibits another catastrophic change, becoming monocentric.

To describe how the analysis works in the nonlinear case, we now provide a detailed analysis of the conditions under which the duocentric configuration is an equilibrium. To this end, consider the configuration depicted in Figure 6.9(a) in which two business sections of equal size are symmetrically located. With

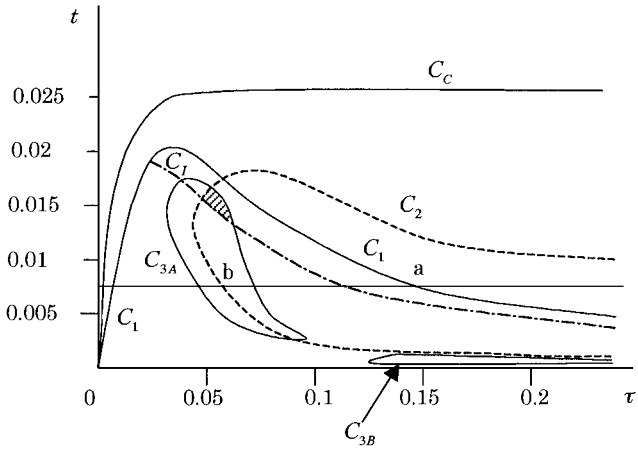


Figure 6.12: Parameter ranges for six equilibrium configurations.

respect to $x \geq 0$, the corresponding density functions are as follows:

$$m(x) = 1/S_{\mathrm{f}} \quad \text{and} \quad n(x) = 0 \quad x \in [b_1, b_3] \tag{6.73}$$
$$m(x) = 0 \quad \text{and} \quad n(x) = 1/S_{\mathrm{h}} \quad x \in [0, b_1] \quad \text{and} \quad [b_3, b_4].$$

Furthermore, the individuals residing in the section $[0, b_1]$ (respectively $[b_3, b_4]$) are assumed to work in the business section $[b_1, b_2]$ (respectively $[b_2, b_3]$). The corresponding commuting function is then

$$\begin{aligned} J(x) &= b_1 + \frac{b_2 - b_1}{b_1}x \quad x \in [0, b_1] \\ &= \frac{b_2 b_4 - b_3^2}{b_4 - b_3} + \frac{b_3 - b_2}{b_4 - b_3}x \quad x \in [b_3, b_4] \\ &= x \quad \text{elsewhere} \end{aligned} \tag{6.74}$$

so that $J(0) = b_1$, $J(b_1) = b_2$, $J(b_3) = b_2$, and $J(b_4) = b_3$. The associated wage function is given by

$$W^*(x) = W^*(b_2) - t|b_2 - x|, \quad x \geq 0, \tag{6.75}$$

which is represented in Figure 6.9(c).

Next, using (6.35), (6.48), and (6.73), we can compute the aggregate accessibility as follows:

$$A^*(x) = \frac{\beta}{S_{\mathrm{h}}}\left[\int_{-b_3}^{-b_1} \exp(-\tau|x - y|)dy + \int_{b_1}^{b_3} \exp(-\tau|x - y|)dy\right], \tag{6.76}$$

which is represented in Figure 6.9(b). Unlike the case depicted in Figure 6.6 for the linear accessibility case, $A^*(x)$ now achieves its maximum at a location inside each business section and decreases as x moves toward 0.

Last, substituting (6.74)–(6.76) into (6.31) and (6.36) leads to the following bid rent functions:

$$\Phi^*(x) \equiv \Phi(x, 0) = \frac{A^*(x) - [W^*(b_2) - t|b_2 - x|]L_f}{S_f} \tag{6.77}$$

$$\Psi^*(x) \equiv \Psi(x, u^*) = \frac{W^*(b_2) - t|b_2 - x| - Z(u^*)}{S_h}. \tag{6.78}$$

Once b_1 has been chosen, all the other boundaries, and hence functions, are uniquely determined by the following expressions:

$$b_2^* = \frac{S_f + S_h L_f}{S_h L_f} b_1 \quad b_3^* = b_1 + \frac{S_f M^*}{2} \quad b_4^* = \frac{(S_f + S_h L_f)M^*}{2}. \tag{6.79}$$

Clearly, symmetry implies that no more than one half of the household population live in the residential section $[0, b_1]$, and thus $b_1 \leq S_h L_f M^*/2$. In fact, the feasible range of b_1 can be further narrowed. Indeed, the land market equilibrium condition requires that the household bid rent at $x = 0$ not be less than R_A and be equal to R_A at the city fringe b_4^*:

$$\Psi^*(0) \geq R_A = \Psi^*(b_4^*),$$

which implies that $b_2^* \leq b_4^*/2$. Substituting into this inequality the first and third equalities in (6.79) yields $b_1 \leq S_h L_f M^*/4$.

For any $b_1 \in [0, S_h L_f M^*/4]$, the corresponding duocentric configuration is a spatial equilibrium if and only if the following conditions hold (see Figure 6.9(d) for an illustration):

$$\Phi^*(x) \leq \Psi^*(x) \quad x \in [0, b_1]$$

$$\Phi^*(x) \geq \Psi^*(x) \quad x \in [b_1, b_3^*]$$

$$\Phi^*(x) \leq \Psi^*(x) \quad x \in [b_3^*, b_4^*]$$

$$\Psi^*(b_4^*) = R_A.$$

Using (1) the fact that the continuous function $A^*(x)$ is strictly concave (convex) on each business section (residential section) and (2) that $W^*(x)$ is linear on each commuting section, the equilibrium conditions above can be restated under the form of conditions evaluated at a finite number of locations:

$$\Phi^*(0) \leq \Psi^*(0) \tag{6.80}$$

$$\Phi^*(b_1) = \Psi^*(b_1) \tag{6.81}$$

$$\Phi^*(b_2^*) \geq \Psi^*(b_2^*) \tag{6.82}$$

$$\Phi^*(b_3^*) = \Psi^*(b_3^*) \tag{6.83}$$

$$\Phi^*(b_4^*) \leq \Psi^*(b_4^*) \tag{6.84}$$

$$\Psi^*(b_4^*) = R_A. \tag{6.85}$$

We may then proceed as follows. First, (6.85) univocally determines $W^*(b_2^*)$ as a function of b_1. In turn, (6.83) uniquely then determines u^*. Second, (6.83) and (6.81) lead to

$$t = \frac{S_\text{h}}{S_\text{f} + S_\text{h} L_\text{f}} \frac{A^*(b_1) - A^*(b_3^*)}{b_1 + b_3^* - 2b_2^*}. \tag{6.86}$$

Third, it follows from (6.81) and (6.80) that

$$t \leq \frac{S_\text{h}}{S_\text{f} + S_\text{h} L_\text{f}} \frac{A^*(b_1) - A^*(0)}{b_1}. \tag{6.87}$$

Fourth, (6.81) and (6.82) yield

$$t \leq \frac{S_\text{h}}{S_\text{f} + S_\text{h} L_\text{f}} \frac{A^*(b_2^*) - A^*(b_1)}{b_2^* - b_1}. \tag{6.88}$$

Finally, from (6.81) and (6.84) we get

$$t \leq \frac{S_\text{h}}{S_\text{f} + S_\text{h} L_\text{f}} \frac{A^*(b_1) - A^*(b_4^*)}{b_1 + b_4^* - 2b_2^*}. \tag{6.89}$$

Consequently, for any $b_1 \in [0, S_\text{h} L_\text{f} M^*/4]$, the associated duocentric configuration is a spatial equilibrium if and only if (6.86), (6.87), (6.88), and (6.89) hold.

The relationship between the three parameters t, τ, and b_1 may be studied numerically by setting the remaining parameters at the values used for Figure 6.12, that is, $(N, S_\text{f}, S_\text{h}, L_\text{f}) = (1000, 1, 0.1, 10)$. We can then draw in the (t, τ)-space the curve corresponding to (6.86). Applying the three remaining conditions (6.87)–(6.89), we can determine the subset of (t, τ)-values for which the duocentric configuration associated with b_1 is an equilibrium. Repeating the same operation by changing b_1 from 0 to $S_\text{h} L_\text{f} M^*/4$ yields the complete set of duocentric equilibria.

For example, when $b_1 = 24$, the associated duocentric configuration is a spatial equilibrium when (t, τ) belongs to the curve passing through points A and B in Figure 6.13; in this case, the upper bound of this curve is determined by (6.87). But for $b_1 = 5$, the associated duocentric configuration is a spatial equilibrium when (t, τ) belongs to the curve passing through points B and C in Figure 6.13; the upper bound of this curve is now determined by (6.88). The union of all these curves defines the domain delimited by the line C_2 in Figure 6.12.

It is worth noting that, for (t, τ) at point B, the duocentric configuration associated with $b_1 = 5$ and $b_1 = 24$ are both an equilibrium. More generally, at every point in the region delineated by C_2 but below the heavy line in Figure 6.13, two duocentric equilibria exist. Thus, we may also have multiplicity of equilibria of the same type. All of this shows the complexity of the analysis with a nonlinear accessibility.

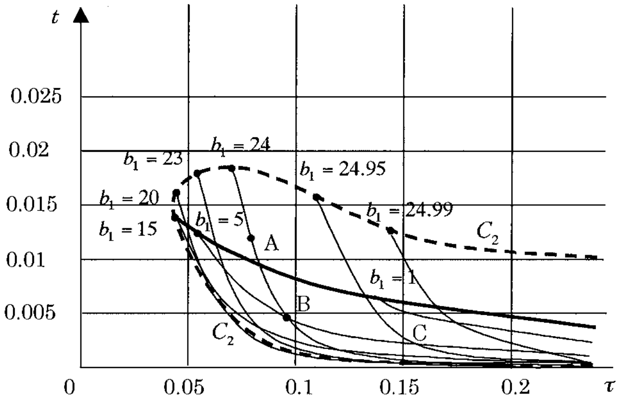


Figure 6.13: The domain of (t, τ) sustaining duocentric equilibria for each admissible value of b_1.

Another important point to make is that diversity in agent types is crucial for the preceding patterns to arise. Considering firms only with the framework of interaction, Tabuchi (1986) obtained results similar to those derived in Section 6.2; that is, the equilibrium distribution of firms is always unimodal because the global accessibility function $A(x)$ remains concave over the whole urban area. On the contrary, with both firms and households, we typically observe an alternation of concavity and convexity. The standard assumption of a single class of agents is, therefore, not as innocuous as might be expected and may well prevent the emergence of more realistic urban patterns.

6.6 SUBURBANIZATION AND THE LOCATION OF MULTIUNIT FIRMS

So far, a firm has been considered as a single-unit entity. Consequently, the model discussed in Section 6.3 is not able to explain a basic trend observed in the spatial organization of large cities, that is, the location of firm units in suburban areas. For example, many firms (e.g., banks or insurance companies, but also industrial firms a long time ago) have moved part of their activities (such as bookkeeping, planning, and employee training) to the suburbs; similar moves had been observed earlier in the case of industrial activities (Hohenberg and Lees 1985, chap. 6). In this case, a firm typically conducts some of its activities (such as communications with other firms) at the front office located in the CBD, whereas the rest of its activities are carried out at the back office set up in the suburbs.

This problem has recently been tackled by Ota and Fujita (1993). With the other assumptions of the model in Section 6.3 kept unchanged, it is now assumed that each firm consists of a front unit and a back unit. Each front unit is assumed to interact with all other front units for business communications, whereas each back unit exchanges information or management services only with the front unit belonging to the same firm. Each firm must choose the location of the front unit and back unit so as to maximize its profit. If a firm sets up its front unit at $x \in X$ and back unit at $z \in X$, the firm incurs an *intrafirm communication cost* $\Gamma(x, z)$ that depends only upon the locations x and z. As before, each front-unit needs S_{f} units of land and L_{f} units of labor; each back unit requires S_{b} units of land and L_{b} units of labor.

In this context, the only change from the previous model is in the profit function (6.34). A firm having a front unit at x, a back unit at z and choosing a level of contact activity $\varphi(x, y)$ with the front unit of each other firm at $y \in X$ has now a profit function defined as follows:

$$\pi(x, z) = A(x) - R(x)S_{\mathrm{f}} - W(x)L_{\mathrm{f}} - R(z)S_{\mathrm{b}} - W(z)L_{\mathrm{b}} - \Gamma(x, z),$$

where the first term is defined by (6.35). If local accessibility is linear (see (6.47)) and the intrafirm communication cost is linear in distance, Ota and Fujita (1993) have shown that no less than 11 different equilibrium configurations are possible, depending on the values of the various parameters. These configurations are the result of two basic effects: (1) as the commuting cost of workers decreases, the segregation of business and residential areas increases, and (2) as the intrafirm communication cost gets smaller, back units separate from front units. The most typical configuration arising when intrafirm communication costs are low involves the agglomeration of the front units at the city center surrounded by a residential area, whereas back units are established at the outskirts of the city together with their employees. In other words, a primary labor market emerges at the city center (e.g., Manhattan), whereas secondary labor markets are created in the far suburbs. Therefore, *the advancement of intrafirm communications technologies provides a major cause for job suburbanization*.[22]

6.7 CONCLUDING REMARKS

According to the conventional wisdom, externalities are at the root of economic agglomerations. And indeed, we have seen in this chapter how nonmarket interactions among economic agents may give rise to different types of agglomerations even in the absence of increasing returns or imperfect competition. In particular, appealing to externalities allows one to save the competitive paradigm in the economics of agglomeration and to use standard models. Thus, the mere social inclinations of people are sufficient for the formation of a settlement exhibiting the main features of a modern city with an endogenous center around which human beings organize themselves. The same type of spatial organization arises among firms benefiting from information spillovers. However, it does not seem easy to get several settlements, districts, or centers with a population of homogeneous agents. At least two groups of agents, such as firms and households, seem necessary for richer spatial patterns to emerge. In addition, nonlinear fields of interaction are required because linear fields tend to support monocentric structures only. Unfortunately, although the use of an exponential distance-decay function is a good approximation of many real-world patterns of interaction, it appears that such models very quickly become hard to manipulate. As is well known, nonlinearities tend to generate multiplicity of equilibria and discontinuous transitions (see, e.g., Grandmont 1988). The contrast between the results obtained under linear and exponential distance-decay functions suggests that the form taken by the social process of interaction in the transmission of knowledge and information is crucial for the type of urban configuration that may emerge.

Not surprisingly, the first theorem of welfare does not apply because of the presence of externalities. Unlike the conventional wisdom, which stresses the negative externalities associated with the formation of cities more, we have singled out here one of the main benefits generated by agglomerations, that is, the spreading of information among close-knit agents. Such an external effect may explain why economic agents are prepared to pay high rents to live close to the centers of large cities where this effect is most intense. Consequently, we have been able to show that, in a market economy, agglomeration of firms or households is desirable from the social point of view. Even less expected, we have seen that, for some configurations of parameters, it is socially optimal to have denser or more agglomerated patterns of agents than those generated by the market. This seems to be a fairly general principle inherent to the process of information dissemination: people account for their role as receivers but not as transmitters. In this case, equilibrium patterns are too spatially dispersed. Surely, this is not what most proponents of spatial planning would expect. Of course, the models discussed in this chapter focus upon specific aspects only, and other agglomeration and dispersion forces must also be studied. In particular, external effects like pollution or crime tend to deter further urban growth and, therefore, favor the dispersion of human activities. In addition, as stated in Section 6.1, the population of households has been assumed to be given. Consequently, our welfare results are valid conditional upon some given population of workers who must reside in the city.

Finally, low commuting costs tend to foster a monocentric urban configuration. This occurs because such low costs allow the nonmarket interactions among firms to become the predominant location factor for firms. Once this fact is understood, it is no surprise that firms want to agglomerate in a single district. This is a general theme that we will encounter again in the next chapter.

APPENDIX

A. Set $v' \equiv dv/dx$ and $v'' \equiv d^2v/dx^2$, where v is a function defined for $x \geq 0$. Given (6.10), we want to solve the following differential equation:

$$v'' = \frac{a}{2}\exp(-v), \tag{A.1}$$

where $v(x) \equiv T(x)/\alpha$ and $a \equiv (\alpha/4t)\exp(\zeta/\alpha)$. Multiplying both sides of (A.1) by v' and integrating, we obtain

$$(v')^2 = -a\exp(-v) + c_1, \tag{A.2}$$

where c_1 is a constant of integration. Set

$$w^2 = -a\exp(-v) + c_1 \tag{A.3}$$

so that

$$2w\,dw = a\exp(-v)dv$$

from which it follows by (A.3) that

$$dv = \frac{2w\,dw}{c_1 - w^2}. \tag{A.4}$$

Denoting $c_1 = k^2$ and using $dv = w\,dx$ in (A.4), we obtain

$$\frac{2dw}{k^2 - w^2} = dx. \tag{A.5}$$

Observe that

$$\frac{2}{k^2 - w^2} = \frac{1}{k}\frac{1}{k+w} + \frac{1}{k}\frac{1}{k-w}.$$

Substituting preceding in the left-hand side of (A.5) and integrating the resulting expression, we have

$$\frac{1}{k}\log\frac{k+w}{k-w} = x + c_2, \tag{A.6}$$

where c_2 is a constant of integration. Solving (A.6) for w gives

$$w = k\frac{\exp k(x+c_2) - 1}{\exp k(x+c_2) + 1}.$$

Using (A.3) and the definition of k, we obtain

$$\frac{k^2}{a} - \exp(-v) = \frac{k^2}{a}\left(\frac{\exp k(|x|+c_2) - 1}{\exp k(|x|+c_2) + 1}\right)^2$$

whose solution is

$$v(x) = -\log\frac{k^2}{a}\left[1 - \left(\frac{\exp k(|x|+c_2) - 1}{\exp k(|x|+c_2) + 1}\right)^2\right]. \tag{A.7}$$

To determine c_2, we observe that $T(x)$ and, therefore, $v(x)$ are minimized at $x = 0$, thus implying that $c_2 = 0$. Consequently, (A.7) may be rewritten as follows:

$$v(x) = -\log \frac{k^2}{a} \frac{4 \exp k|x|}{(1+\exp k|x|)^2},$$

which is equivalent to

$$T(x) = -\alpha \ \log \left[\frac{\alpha}{t} \exp\left(-\frac{\zeta}{\alpha}\right) \frac{k^2 \exp(k|x|)}{(1+\exp(k|x|))^2} \right]$$

after having replaced a by its value.

B. We derive here the optimality conditions of Section 6.2.1 by expressing the optimization problem as an optimal control problem. However, this process turns out to be fairly involved because trips between locations go both ways. It is worth noting that we encounter here one of the main differences between time and space modeling, for interactions in space are essentially bidirectional.[23]

Differentiating (6.3), we obtain

$$\begin{aligned} \frac{dT}{dx} &= \int_{-b}^{x} tn(y)dy - \int_{x}^{b} tn(y)dy \\ &= tN(x) - t[N - N(x)] \\ &= 2tN(x) - tN, \end{aligned}$$

where

$$N(x) \equiv \int_{-b}^{x} n(y)dy$$

denotes the total population situated on the left of x. By definition of $N(x)$, we have

$$\frac{dN}{dx} = n(x)$$

together with the terminal conditions $N(-b) = 0$ and $N(b) = N$.

Specifying the initial condition on $T(x)$ appears to be especially complex here because the value of $T(-b)$ depends on the entire population distribution:

$$T(-b) = \Gamma(-b) + tbN,$$

where

$$\Gamma(x) \equiv \int_{x}^{b} tyn(y)dy,$$

and thus we have the new differential equation:

$$\frac{d\Gamma}{dx} = -txn(x)$$

with the corresponding terminal condition $\Gamma(b) = 0$.

Consequently, our optimization may now be rewritten as a standard optimal control problem under the form:

$$\max_{b,n(\cdot)} S = \int_{-b}^{b} \{[Y - U^* - \alpha \log n(x) + I - T(x)]n(x) - R_A\} dx$$

subject to the constraints

$$\frac{dT(x)}{dx} = 2tN(x) - tN$$
$$\frac{dN(x)}{dx} = n(x)$$
$$\frac{d\Gamma(x)}{dx} = -txn(x)$$

together with the terminal conditions

$$T(-b) = \Gamma(-b) + tbN \quad N(-b) = 0 \quad \text{and} \quad N(b) = N \quad \Gamma(b) = 0.$$

As a result, we have the following Hamiltonian:

$$H(x) = [Y - U^* - \alpha \log n(x) + I - T(x)]n(x) - R_A$$
$$+ \lambda(x)[2tN(x) - tN] + \mu(x)n(x) - \nu(x)txn(x),$$

where $\lambda(x)$, $\mu(x)$, and $\nu(x)$ and the costate variables (multipliers) are associated respectively with the state variables $T(x)$, $N(x)$, and $\Gamma(x)$. By solving the motion equations of the costate variables, we get

$$\lambda(x) = N(x) - N$$
$$\mu(x) = -T(x) + txN + T(-b) + \mu(-b)$$
$$\nu(x) = N.$$

Setting

$$\zeta^o \equiv Y - U^* + I - \alpha + T(-b) + \mu(-b)$$

and substituting the costate variables by their values, the Hamiltonian may be rewritten as

$$H(x) = [\alpha + \zeta^o - \alpha \ \log n(x) - 2T(x)]n(x) - R_A$$
$$+ [N(x) - N][2tN(x) - tN]).$$

Applying the first-order condition with respect to n leads to the solution (6.22).

Finally, because b is chosen without constraint, the Hamiltonian evaluated at b must be equal to zero, which means that (6.24) holds since $N(b) = N$.

C. By the land constraint (6.41), we must have $m(y) \leq 1/S_f$ for all $y \in X$. Furthermore, the firm population constraint (6.44) implies that

$$\begin{aligned} A(x) &= \int_{-\infty}^{\infty} m(y) \exp(-\tau|x-y|)dy \\ &< \int_{-\infty}^{\infty} \frac{1}{S_f} \exp(-\tau|x-y|)dy \\ &= \frac{2}{S_f} \int_{0}^{\infty} \exp(-\tau z)dz \\ &= \frac{2}{\tau S_f}, \end{aligned}$$

which gives us the desired inequality.

NOTES

1. Even before the Industrial Revolution, the exchange of information seems to have been the main cause for the emergence of a prominent center in the business community (think of Venice, Antwerp, and Amsterdam) in different time periods (Smith 1984).
2. In Section 3.2, recall that we studied the location of activities but not that of firms.
3. Hotelling's idea to model spatial markets using a continuous distribution was intended to make sure that individual discontinuities, due to nonconvex preferences, would be distributed in such a way that they would not be noticeable to any firm, thus making each firm's demand continuous.
4. Such an approach to city formation agrees with Glaeser, Kolko, and Saiz (2001) for whom cities hinge more and more on consumption amenities.
5. In the same vein, Papageorgiou and Smith (1983) consider a trade-off between the need for social contacts, which is negatively affected by distance, and the need for land, which is negatively affected by crowding. Space is circular (a circle or a taurus) so that the uniform distribution of individuals is always a spatial equilibrium. When the propensity to interact with others is large enough, this equilibrium becomes unstable: any marginal perturbation is sufficient for the population to evolve toward an irregular distribution. In this model, cities are considered as the outcome of a social process combining basic human needs, which are not (necessarily) expressed through the market. It is probably fair to say that this model captures much of the intuition of early geographers interested in the spatial structure of human settlements in that the key variables are independent of the economic system. However, from the economic standpoint, it is important to consider less abstract formulations and, rather, to study models based on specific economic interactions.
6. To be precise, we should say any nonzero measure set of households.
7. A formal derivation of these conditions is provided in Part B of the chapter appendix.
8. See Lucas (2001) for the study of a similar problem in a more general context using a two-dimensional space and a general production function.

9. Because land and firms are assumed to be owned by absentee landlords and shareholders, the income of each consumer equals her wage earned at the chosen job site.
10. In other words, no firm refuses a contact initiated by any other firm. This seems natural because of the symmetry of the communication environment considered.
11. Observe that communications externalities may find their origin in previous innovations occurring in the same production sector or in innovations made in other sectors. Thus, they are consistent with both localization and urbanization economies.
12. Note that (6.47) arises when we set $c_1(x, y) = [\log(1 - \tau|x - y|)]^{-1}$, which is an increasing and convex function of $|x - y|$.
13. Observe that the equilibrium wage rate is not unique in the interval $[b_1^*, b_2^*]$ because no firm is set up there. However, expression (6.58) is sufficient to show that no firm wants to locate in this interval.
14. As usual, one must also check that $\max\{\Phi^*(x), \Psi^*(x)\} \geq R_A$ for $x \in [0, b_2^*]$ while $\max\{\Phi^*(x), \Psi^*(x)\} < R_A$ for $x > b_2^*$.
15. The same type of externality has been further explored by Kanemoto (1990), who considered the case in which firms are engaged in transactions with others. Combining the exchange of intermediate inputs between firms with indivisibilities in their production creates externalities similar to those considered by Fujita–Imai–Ogawa. If τ is the unit transportation cost of the intermediate goods, Kanemoto then showed that the monocentric configuration is an equilibrium when the ratio t/τ is small, which is a condition similar to Proposition 6.3.
16. See Proposition 3.2 for a similar result.
17. For an extension of the present study to a two-dimensional space, see Ogawa and Fujita (1989).
18. For details, see Fujita (1985).
19. For simplicity, we neglect the inefficiency arising from the cost c_2 imposed by a firm on another when getting information from it.
20. Observe that the same holds when the city size is small enough for the linear approximation to be acceptable.
21. It is worth noting that heuristic models of urban land use developed by some geographers yield pattern displaying similar phenomena (see, e.g., Allen and Sanglier 1979; 1981).
22. Note that the delocation of production plants from industrialized countries toward developing countries where labor is cheaper can be given a similar explanation, although other factors, more in line with the neoclassical trade theory, are also involved.
23. Formally, this means that we have to solve a problem of the calculus of variations with multiple integrals.

7

The Formation of Urban Centers under Imperfect Competition

7.1 INTRODUCTION

The analysis developed in the previous chapter sheds light on the emergence of centers within a city. There, the emergence of an urban center crucially depends on the existence of nonmarket interactions (externalities) among agents and is typical of the formation of a CBD involving high-level activities. However, in the real world, we also frequently observe the formation of *clusters of stores* selling similar goods (fashion clothes, restaurants, movie theaters, antiquity shops, etc.) or of *employment centers* in which different kinds of jobs are performed. In such cases, the agglomeration forces are created through market interactions between firms and consumers or workers. As seen in Chapter 2, for this to occur, one must consider increasing returns and imperfect competition. In this chapter, our purpose is to show how urban centers of different types, such as commercial areas or employment centers, may emerge under imperfect competition in the product or labor market. We will consider different market structures. The common thread of the various models considered here is that monopoly (or monopsony) power on the product (labor) market is needed for such agglomerations to emerge in equilibrium.

In Section 7.2, we suppose monopolistic competition with a continuum of firms selling differentiated varieties (Fujita 1986; 1988). Firms no longer assume that they can sell whatever they want at given market prices. Instead, each firm is aware that its optimal choice (location and price) depends on the demand for the variety it supplies. This demand itself rests on the spatial consumer distribution, thus showing how firms' choices are directly affected by consumers' choices. In turn, the optimal choice of a consumer (location and consumption) depends on the entire firm distribution. This is so because firms sell a differentiated product and consumers like variety; hence, their purchases are distributed across locations and the distribution of their shopping trips varies with the number of varieties available at each location. This creates some form of spatial interdependence between the two distributions that is solved through the interplay between the firms' and consumers' bid rent functions. It is interesting to observe that the knowledge of prices alone does not allow firms and households to make their optimal choices. As noted by Koopmans (1957, 154) in a different, but related, context,

> The decisive difficulty is that transportation of intermediate commodities from one plant to another makes the relative advantage of a given location for a given plant dependent on the locations of other plants.

The same observation applies to consumption goods traded between firms and households (Papageorgiou and Thisse 1985). The most typical outcome of this kind of interaction involves a district in which all firms are agglomerated together with some consumers and surrounded by two residential sections. That firms sell differentiated varieties and therefore price above marginal cost allows them to compete for land with the aim of being close to consumers. In the absence of such a positive markup, firms would lose their incentives to get close to their customers and the agglomeration would vanish. This analysis shows how imperfect competition operates to generate agglomeration through market transactions and also sheds additional light on the meaning and implication of the spatial impossibility theorem discussed in Chapter 2.

The monopolistic competition model is easy to handle but fails to capture one of the basic ingredients of spatial competition, that is, the strategic nature of spatial competition, as discussed in Chapter 4. This problem was addressed by Hotelling (1929) in the special case of a homogeneous product, and since then it has been generally accepted that competition for market areas is a centripetal force that will lead vendors to congregate – a result known in the literature as the principle of minimum differentiation. This principle has generated controversies about the inefficiency of free competition, for, according to Hotelling (1929, 54) himself, it would suggest that "buyers are confronted everywhere with an excessive sameness."

The classical two ice-cream men problem provides a neat illustration of this principle. Two merchants selling the same ice cream at a the same fixed price compete in location for consumers who are uniformly distributed along a bounded linear segment (Main Street). Each consumer purchases one ice cream from the nearer seller. The consumers are thus divided into two segments, and each firm's aggregate demand is represented by the length of its market segment. Since Lerner and Singer (1937), it has been well known that the unique Nash equilibrium of this game is given by the two firms located at the market center, and so regardless of the shape of the transport cost function. In other words, two firms competing for clients choose to minimize their spatial differentiation. However, matters become more complex when (mill) prices are brought into the picture because price competition is so fierce under product homogeneity that firms always want to separate in order to benefit from the monopoly power generated by geographical isolation.[1]

Spatial separation may cease to be profitable, however, when firms sell a differentiated product. This is so because demand for a firm's variety now arises at each and every consumer location. In addition, it is a well-established result in industrial organization that product differentiation relaxes price competition. It therefore becomes natural to investigate the possibility of an agglomeration of oligopolistic firms when there is a strong preference for variety.

Using a framework similar to the one in the preceding section, we show in Section 7.3.1 that, despite the presence of strategic price competition, the agglomeration of firms at the market center is a Nash equilibrium of the corresponding noncooperative game when firms supply sufficiently differentiated varieties, assuming that the consumer distribution is given. By putting together all these results, we may conclude that, regardless of the market structure, *a commercial area involving a large number of stores, restaurants, or theaters is likely to emerge when it offers sufficiently differentiated products, when the transport cost borne by consumers are low enough, or both.* Although the conventional wisdom sees such clusters as socially wasteful, we will show in Section 7.3.2 that they are often socially desirable once their ability to supply differentiated goods is taken into account.

The foregoing models can be referred to as *shopping* models because consumers visit firms and bear the entire transportation costs. Instead, we have *shipping* models when firms deliver the product and take advantage of the fact that the customers' locations are observable to price discriminate across locations. Shopping models seem to be appropriate for studying competition among sellers of consumption goods, whereas shipping models better describe competition among sellers of industrial goods. However, the possibility of ordering through such communications technologies as the telephone and Internet and the existence of mail-order firms make these settings increasingly relevant for the study of consumption goods too. Despite significant differences in the process of competition, the tendency toward agglomeration will be shown in Section 7.3.3 to be governed by principles similar to those uncovered for shopping models. As will be seen, strategic interaction is at the heart of these two families of models, and space is the reason for it: competition is localized in shopping models while shipping models involve oligopolistic competition in spatially separated markets. Finally, we conclude in Section 7.3.4 with a brief survey of what has been accomplished in spatial competition when consumers make multipurpose trips.

Once it is recognized that consumers shop for the best price and variety opportunity on the purchasing day, it becomes sensible to assume that they have incomplete information about which firm offers which variety at which price. Although search theory has experienced a rapid growth since the early 1970s (McMillan and Rothschild 1994), it is fair to say that space brings about specific dimensions that have attracted the attention of only a few analysts. Indeed, because trading arises in a number of places much smaller than where consumers reside, they experience different search costs, and the way consumers conduct their search affects firms' strategies through their demand. This in turn implies that search costs are influenced by firms' strategic decisions such as locations and prices. In such a context of comparison shopping, the clustering of shops is based on consumers' economizing on their search cost (Nelson 1970). More precisely, Stuart (1979, 19) has noted that "a seller who does locate as a spatial monopolist might have a hard time attracting search-conscious buyers in the first place." This observation has a major implication for our purpose in that "a spatial clustering of sellers can result from desires of buyers to search in marketplaces where there are relatively many sellers" (Stuart 1979, 17).

Indeed, consumers unaware of the characteristics of the varieties supplied in various places reduce their search costs by visiting the place with the largest number of stores even though this place is located farther away. Hence, incomplete information on the consumer side is an agglomeration force. This problem has been tackled by different authors (Eaton and Lipsey 1979; Stahl 1982; Wolinsky 1983; Schulz and Stahl 1996). Common to all these contributions is that the expected utility from visiting a cluster of firms increases with its size, which is a result reminiscent of the gravity principle. Although each consumer buys a single variety, in the aggregate consumers exhibit a preference for variety because of their lack of information about the available varieties. Furthermore, consumers are affected differentially according to their distances to the marketplaces. In Section 7.4, we show how the agglomeration force is generated from the aggregate behavior of individual consumers pursuing a search strategy. We will first focus on a standard model of spatial competition in which prices are given. We briefly discuss what happens when the analysis is then extended to deal with price competition and variable total demand.

In all cases, if consumers have different tastes and are uncertain about the characteristics of the varieties on offer, the firms can manipulate the search cost structure by joining an existing market or by establishing a new one. The basic trade-off faced by a firm is as follows: a firm captures a small market share when setting up in a large market or monopolizes a small local market when opening a new one. When a firm chooses to join the cluster, it generates a *demand externality* in that more consumers will benefit from economies of scope in searching, thus increasing the number of consumers visiting the cluster (that is, the extent of the product market is endogenous). Such an externality is obviously a centripetal force similar to the network externalities encountered in the consumption of goods whose utility increases with the number of users (e.g., telephone, e-mail, etc.). This externality is also akin to the agglomeration force that we will encounter in Chapter 9.

After having studied the formation of commercial areas, it is natural to move to the creation of employment centers in Section 7.5. As we know, the CBD is a natural place in which employment may be concentrated. However, contrary to general beliefs, the suburbanization of jobs is a not a new phenomenon.

As noticed by Hohenberg and Lees (1985, 131), it arose, for example, in protoindustrial Europe:

> Big-city entrepreneurs took advantage of lower rural wages and of heightened division of labor in decentralizing parts of their production while reserving the more delicate operations to city artisans.

Today, the creation of suburbanized jobs seems to obey to a similar logic, although it may take different forms such as the emergence of *edge cities* (Henderson and Mitra 1996). Thus, our research strategy will be similar to that developed in the previous sections. We first assume that employers have no market power on the labor market (Section 7.5.1) and go on by analyzing the case of a firm having some monopsony power on the urban labor market (Section 7.5.2). From the formal point of view, the results obtained in this section can be viewed as the dual of those obtained in the previous ones.

Although the models presented in this chapter may look very different from one another, they are in essence analogous and address similar questions. In particular, they all deal with the formation of shopping areas as well as with the dual concept of employment subcenters.

7.2 MONOPOLISTIC COMPETITION AND THE FORMATION OF SHOPPING DISTRICTS

Consider an economy with two types of goods. The first one is homogeneous; it is supplied on a perfectly competitive market and serves as the numéraire. The second good is a horizontally differentiated product produced under increasing returns to scale and imperfect competition. Each firm in this sector has a negligible impact on the market outcome in the sense that it can ignore its impact on, and hence reactions from, other firms. To this end, we assume that there is a continuum M of firms.[2] In addition, each firm sells a differentiated variety and therefore faces a downward sloping demand. Consequently, our model is one of monopolistic competition in which all the unknowns are described by density functions. There are no scope economies, and thus, owing to increasing returns to scale, there is a one-to-one relationship between firms and varieties. Each firm faces the same technology; it uses the fixed amount of land S_f and bears a fixed cost f and a constant marginal cost c expressed in terms of the numéraire. Finally, firms are owned by absentee shareholders.

Space is linear and given by $X = (-\infty, \infty)$. The amount of land at each location is equal to 1, and land is owned by absentee landlords. Each variety can be traded at a positive cost of t units of the numéraire for each unit transported over one unit of distance, regardless of the variety. In other words, transportation costs are linear in distance and quantity.

There is a continuum N of consumers, each using the same fixed amount of land S_h. Preferences are identical and are described by the following additive utility function, which is symmetric in all varieties:

$$U(z; q(i), i \in [0, M]) = \int_0^M u[q(i)]di + z, \tag{7.1}$$

where u is strictly concave and increasing until the level of satiation is reached, $q(i)$ the quantity of variety $i \in [0, M]$, and z the quantity of the numéraire. There is no multipurpose trip, and each unit of each variety is bought on a single trip (think of restaurants, theaters, etc.). If the firm supplying variety i locates at $y(i) \in X$, the budget constraint of a consumer located at $x \in X$ can be written as follows:

$$\int_0^M [p(i) + t|x - y(i)|]q(i)di + R(x)S_h + z = Y,$$

where Y is the consumer's income which is given and the same across consumers, $p(i)$ the price of variety i, and $R(x)$ the land rent at location x. The income Y is supposed to be large enough for the optimal consumption of the numéraire to be strictly positive for each individual.

Because of identical production functions, transportation costs, and symmetric preferences, in equilibrium all varieties provided at the same location y must be supplied at the same (mill) price $p(y)$. Hence, owing to the concavity of u, the quantity of each variety purchased at location y by any consumer at x is the same across varieties:

$$q[i; x; y(i) = y] = q(x, y) \quad \text{for each variety available at } y.$$

Then, if $m(x)$ is the density of firms at $y \in X$, the indirect utility of a consumer at x is

$$V(x) = \int_X u[q(x, y)]m(y)dy - \int_X [p(y) + t|x - y|]q(x, y)m(y)dy \\ - R(x)S_h + Y. \tag{7.2}$$

Regarding the supply side, if the number (formally, the density) of consumers at x is $n(x)$, the profit made by a firm located at x and facing its demand field is

$$\pi(y) = [p(y) - c]\int_X q(x, y)n(x)dx - R(y)S_f - f. \tag{7.3}$$

Given the firm and price densities $m(\cdot)$ and $p(\cdot)$ as well as the land rent $R(\cdot)$, each consumer chooses a location x and demand distribution $q(x, \cdot)$ so as to maximize his indirect utility (7.2). Given the densities $n(\cdot)$, $q(\cdot, \cdot)$, and $R(\cdot)$, each firm selects a location y and its price $p(y)$ to maximize its profits (7.3). As usual, the equilibrium is such that all consumers reach the same utility level across occupied locations and firms earn the same profits at each occupied location. The opportunity cost of land is given by R_A.

7.2.1 Spatial Equilibrium

To derive some simple results and facilitate the comparison with the oligopolistic case dealt with in the next section, we suppose that the utility (7.1) has an entropy-type form:

$$u(q) = \frac{q}{\alpha}(1 + \log\beta) - \frac{q}{\alpha}\log\frac{q}{\alpha} \quad \text{if } q < \alpha\beta \tag{7.4}$$
$$= \beta, \qquad \text{if } q \geq \alpha\beta$$

where α and β are two positive constants such that u is strictly concave up to $q = \alpha\beta$. Although the utility (7.1) is additive, we will see below that the parameter α can be interpreted as an inverse measure of the degree of differentiation between varieties: the higher α is, the less differentiated are the varieties. It can also be readily verified that the utility level strictly increases with M, thus implying that the utility (7.1) exhibits a preference for variety. Finally, the higher β is, the higher the level of satiation in each variety.[3]

Plugging (7.4) into (7.2) and maximizing the resulting expression with respect to $q(x, y)$ for each location pair (x, y), we obtain

$$q^*(x, y) = \alpha\beta \exp -\alpha[p(y) + t|x - y|] \quad x, y \in X. \tag{7.5}$$

Hence, the demand by a consumer at x for a variety supplied at y is described by the exponential distance-decay function considered in the previous chapter (when $\tau = \alpha t$). The main difference is that now the firm's price enters as a variable. This demand describes the "interaction" between a consumer at x and a firm at y, but the interaction now goes through the market because it results from the choices made by both firm and consumer in the market. Observe that (7.5) is independent of the prices charged by the other firms, and thus each firm may behave like a monopolist in selecting its price.

Plugging (7.5) into (7.3) and maximizing the resulting expression with respect to $p(y)$, we obtain the firm's profit-maximizing price under monopolistic competition:

$$p^* \equiv p^*(y) = c + 1/\alpha. \tag{7.6}$$

In words, the equilibrium (mill) price is equal to the marginal cost plus a markup that increases with $1/\alpha$; clearly the equilibrium converges toward marginal cost when $\alpha \to \infty$. This means that α plays exactly the role of a substitution parameter between varieties, and will be interpreted in this way, although there is no direct substitution among varieties. Another way to say the same thing is that the price elasticity of (7.5) is equal to $\alpha p(y)$. Accordingly, a larger value for α implies a more elastic demand for each variety *as if* they were closer substitutes.

Note that the price (7.6) is independent of the distributions of firms and households; it is also independent of the firm's location and of the transportation rate t. The reasons for these strong properties are to be found in the very simple model considered here. They will be relaxed in subsequent developments.

Without loss of generality, we may choose the units of M and N such that $S_{\text{f}} = S_{\text{h}} = 1$. This means that if the actual number of firms is smaller than the actual number of households but if firms use much more land than households ($S_{\text{f}} \gg S_{\text{h}}$), we may well have $M > N$. Replacing $p(y)$ by p^* throughout and using (7.5), we obtain:

$$V(x) = \gamma \int_X m(y) \exp(-\alpha t|x - y|)\, dy - R(x) + Y \tag{7.7}$$

$$\pi(y) = \gamma \int_X n(x) \exp(-\alpha t|x - y|)\, dx - R(y) - f, \tag{7.8}$$

where $\gamma \equiv \beta \exp -(\alpha c + 1)$.

These two expressions are very similar to the profit function obtained in Section 6.3 under spatially discounted accessibility. The fundamental difference is that here the "interaction" takes place between agents belonging to different groups, whereas in Section 6.3 interaction develops among firms only. This implies that *firms are attracted by consumers* and, likewise, that *consumers are attracted by firms*. However, because of land competition, *firms are repulsed by firms and consumers by consumers*. In a nutshell, this says that the agents of a given group are attracted by those of the other group but repulsed by those of the same group. This mutual attraction of firms and households is the agglomeration force, whereas competition for land is the dispersion force.

The behavior of the two groups of agents is perfectly symmetric except that, in (7.7), the last two terms ($Y - R(x)$) would represent an income varying with the consumer's location, whereas the same terms in (7.8) would correspond to a fixed cost ($f + R(x)$) changing with the firm's location.

As usual, we define the households' and firms' bid rent functions as follows:

$$\Psi(x, U^*) = \gamma \int_X m(y) \exp(-\alpha t|x - y|)\, dy + Y - U^* \tag{7.9}$$

$$\Phi(y, \pi^*) = \gamma \int_X n(x) \exp(-\alpha t|x - y|)\, dx - f - \pi^*. \tag{7.10}$$

The unknowns being $m^*(x)$, $n^*(x)$, $R^*(x)$, U^*, and π^*, the equilibrium conditions are straightforward and may be written as

$$R^*(x) = \max\{\Psi(x, U^*), \Phi(x, \pi^*), R_{\text{A}}\} \tag{7.11}$$

$$\Psi(x, U^*) = R^*(x) \quad \text{if } n^*(x) > 0 \tag{7.12}$$

$$\Phi(x, \pi^*) = R^*(x) \quad \text{if } m^*(x) > 0 \tag{7.13}$$

$$n^*(x) + m^*(x) = 1 \quad \text{if } R^*(x) > R_{\text{A}} \tag{7.14}$$

$$\int_X m^*(y) dy = M \tag{7.15}$$

$$\int_X n^*(y) dy = N. \tag{7.16}$$

The following properties are shown in Part A of the chapter appendix: the bid rent curve $\Psi(x, U^*)$ is strictly concave on any business section ($m(x) = 1$) and strictly convex on any residential section ($n(x) = 1$); $\Phi(x, \pi^*)$ is strictly convex on any business section and strictly concave on any residential section. This prevents the emergence of any specialized section surrounded by two areas in which agents of the other group are (exclusively or partially) located. Suppose, indeed, that $[b, b']$ is a business area surrounded by residential or integrated sections, or both, thus implying that $\Phi(b, \pi^*) = \Psi(b, U^*)$ and $\Phi(b', \pi^*) = \Psi(b', U^*)$. Because $\Psi(x, U^*)$ is strictly concave and $\Phi(x, \pi^*)$ is strictly convex over this area, we must have $\Phi(x, \pi^*) < \Psi(x, U^*)$ for all $b < x < b'$, which contradicts the equilibrium condition $\Phi(x, \pi^*) \geq \Psi(x, U^*)$ for all $b < x < b'$. The same argument applies if $[b, b']$ is a residential area surrounded by two business or integrated sections, or both. Using a similar argument, one can readily show that there is no vacant land inside the city.

Consequently, focusing on symmetric patterns,[4] we conclude that the city must involve a single integrated district surrounded by two residential or business sections. A centrally integrated district is of course a consequence of the agglomeration force generated by the mutual attraction of firms and households. Only two configurations are then possible: (1) all firms are located with some consumers in the central district surrounded by two residential sections, or (2) all consumers reside within the central district together with some firms whereas the remaining firms occupy the two adjacent areas. These two equilibrium patterns are depicted in Figure 7.1.

Consider first the case in which all firms, together with some consumers, locate in the central district $[-b_0, b_0]$ surrounded by two residential sections $[-b_1, -b_0]$ and $[b_0, b_1]$. We begin by showing that both densities are constant over $[-b_0, b_0]$. Because the district is integrated, it follows that $\Phi(x, \pi^*) = \Psi(x, U^*)$ for all $x \in [-b_0, b_0]$. Then, using (7.9) and (7.10), we obtain

$$\int_X [m^*(y) - n^*(y)] \exp(-\alpha t|x - y|)\, dy = k, \quad x \in [-b_0, b_0],$$

where k is an unknown constant equal to $-(Y - U^* + f + \pi^*)/\gamma$. This expression can be rewritten as follows:

$$\begin{aligned}\int_{-b_0}^{x} &[m^*(y) - n^*(y)] \exp[-\alpha t(x - y)]dy \\ &+ \int_{x}^{b_0} [m^*(y) - n^*(y)] \exp[-\alpha t(y - x)]dy \\ &= k + \int_{-b_1}^{-b_0} \exp[-\alpha t(x - y)]dy + \int_{b_0}^{b_1} \exp[-\alpha t(y - x)]dy.\end{aligned} \tag{7.17}$$

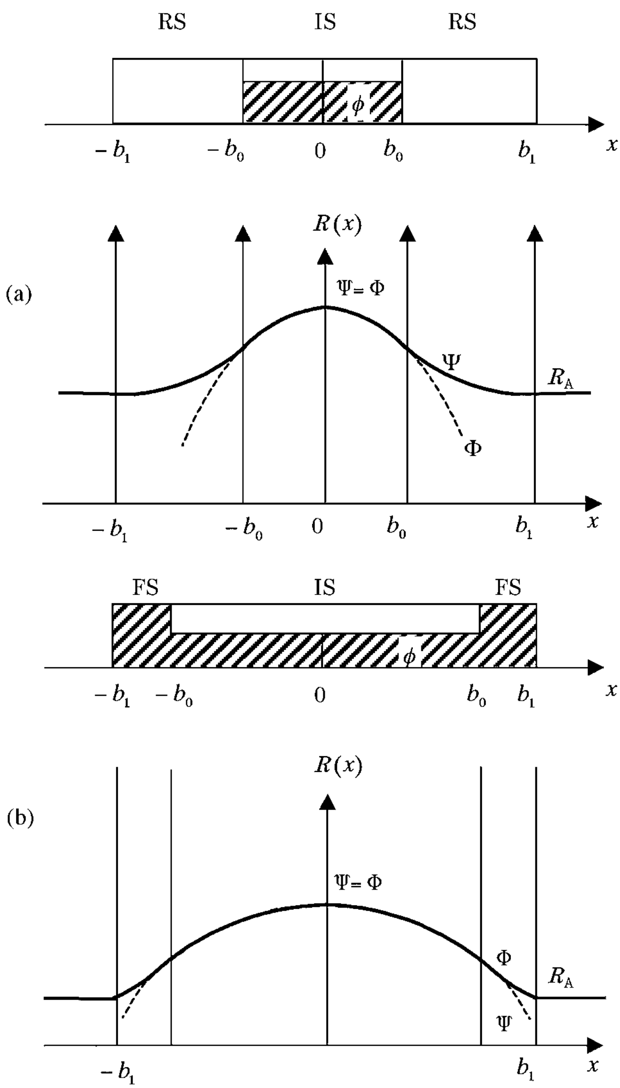


Figure 7.1: The equilibrium city configurations when consumers love variety.

Differentiating (7.17) with respect to x yields

$$\int_{-b_0}^{x} [m^*(y) - n^*(y)] \exp[-\alpha t(x-y)]dy - \int_{x}^{b_0} [m^*(y) - n^*(y)] \exp[-\alpha t(y-x)]dy = \int_{-b_1}^{-b_0} \exp[-\alpha t(x-y)]dy - \int_{b_0}^{b_1} \exp[-\alpha t(y-x)]dy. \quad (7.18)$$

Differentiating again yields

$$2[m^*(x) - n^*(x)] - \alpha t \int_{-b_0}^{b_0} [m^*(y) - n^*(y)] \exp(-\alpha t|x-y|)\, dy$$
$$= -\alpha t \int_{-b_1}^{-b_0} \exp[-\alpha t(x-y)]dy - \alpha t \int_{b_0}^{b_1} \exp[-\alpha t(y-x)]dy. \quad (7.19)$$

Multiplying (7.17) by αt and adding (7.19), we get

$$m^*(x) - n^*(x) = \alpha t k/2 \quad x \in [-b_0, b_0]. \quad (7.20)$$

This, together with the land constraint $m(x) + n(x) = 1$ for $x \in [-b_0, b_0]$, implies that

$$m^*(x) = (1 + \alpha t k/2)/2 \equiv \phi \quad x \in [-b_0, b_0]$$
$$n^*(x) = 1 - m^*(x) = (1 - \alpha t k/2)/2 \quad x \in [-b_0, b_0].$$

It remains to determine the value of ϕ, which gives us that of k. To this end, we add (7.17) and (7.18) and substitute (7.20) into the resulting expression:

$$\alpha t k \int_{-b_0}^{x} \exp[-\alpha t(x-y)]dy$$
$$= k + 2\int_{-b_1}^{-b_0} \exp[-\alpha t(x-y)]dy \quad x \in [-b_0, b_0].$$

Integrating yields

$$\exp(-\alpha t x)[\alpha t k \exp(-\alpha t b_0) + 2\exp(-\alpha t b_0) - 2\exp(-\alpha t b_1)]$$
$$= 0, \quad x \in [-b_0, b_0],$$

which holds provided that

$$\alpha t k \exp(-\alpha t b_0) + 2\exp(-\alpha t b_0) - 2\exp(-\alpha t b_1) = 0,$$

or

$$\exp \alpha t(b_1 - b_0) = 1/2\phi. \quad (7.21)$$

Therefore, we must determine b_0^*, b_1^*, and ϕ simultaneously. The population constraints imply that

$$M = 2b_0 m^*(x) = 2b_0^*\phi \quad (7.22)$$
$$N = 2b_0^* n^*(x) + 2(b_1^* - b_0^*) = 2b_0^*(1-\phi) + 2(b_1^* - b_0^*). \quad (7.23)$$

Taking the logarithm of (7.21) and solving for $b_1^* - b_0^*$, solving (7.22) for b_0^*, and substituting the results in (7.23), we obtain

$$-\frac{2}{\alpha t}\log 2\phi = N - \frac{1-\phi}{\phi}M,$$

which is equivalent to

$$\frac{M}{\phi} = M + N + \frac{2}{\alpha t} \log 2\phi. \tag{7.24}$$

Because $b_1^* - b_0^* > 0$, it follows from (7.21) that ϕ must be positive and smaller than $1/2$. It can then readily be verified that (7.24) has a single solution less than $1/2$ if and only if $N > M$, that is, there are more consumers than firms. Once ϕ is determined, one may obtain k, b_0^*, and b_1^*, and hence U^* and π^*.

When $N = M$, $\phi = 1/2$ is the only solution of (7.24), which means that all firms and households are integrated in a single district.

Finally, it is easy to see that the case in which all consumers are located in the central district $[-b_0, b_0]$ surrounded by the business sections is obtained when $M > N$.

To sum up, we have shown the following:

Proposition 7.1 *There always exists a single spatial equilibrium. This equilibrium is such that*

1. *When $M < N$, all firms are concentrated into a central district with a constant density less than 1/2, land being shared with some consumers, whereas the remaining consumers reside in the two adjacent areas.*
2. *When $M > N$, all consumers are concentrated into a central district with a constant density less than 1/2, land being shared with some firms, whereas the remaining firms are located in the two surrounding sections.*
3. *When $M = N$, the city is formed by an integrated district where the common density is 1/2.*

Hence, we see that land around the middle of the city is used by both groups of agents.[5] This is not unrealistic because one often observes some residences mixed with commercial activities (e.g., Soho in New York City, Montparnasse in Paris, or Shinjuku in Tokyo).

As stated above, both firms and households want to share the central area because of their mutual attraction. Firms do not exclusively use the whole central district, for otherwise firms near $x = 0$ would be (much) less attractive than firms situated at the district fringes $x = -b$ and $x = b$ because of the sharp decrease in their own demand (see (7.5)).[6] This effect is strengthened because the region around $x = 0$ is very attractive for consumers, for they like to distribute their purchase among firms (preference for variety), thus endowing these locations with the highest accessibility to symmetrically distributed stores.

The share of each group inside the integrated district, however, varies with the key parameters α and t. Specifically, it can be shown from (7.24) that the density ϕ of firms decreases as α or t rises. In other words, the packing of firms is less dense when varieties are less differentiated, when transportation is more expensive, or both.[7] The reason is that, in both cases, the demand field $q^*(x, y)$ falls more sharply with distance, inducing firms to get closer to some segments of consumers. In particular, when varieties are "very close substitutes," the equilibrium price is slightly above marginal cost and the land rent is almost flat and close to R_A. Consumers now buy very little from firms not located in their close vicinity, and thus total transportation costs in the economy are also very low. In the limit, when $\alpha \to \infty$, consumers buy only from firms located at the same place as them (backyard capitalism) or refrain from buying the differentiated product.

7.2.2 The Efficiency of the Spatial Equilibrium

We now move to the characterization of the first best optimum when there are N consumers and M firms. As usual, we consider the optimal solution in which the utility level U^* is achieved for the N consumers through minimizing social cost. The decision variables are the consumer density $n(x)$, the firm density $m(x)$, the demand density of a consumer at $x \in X$ for the differentiated product $q(x, .)$, and the consumption of the numéraire by a consumer $z(x)$ at each x. Then, the corresponding total cost is

$$C = \int_X \left[\int_X q(x, y)\,(c + t|x - y|)\, m\,(y) dy \right] n(x) dx + \int_X z(x) n(x) dx \\ + R_A(M + N) + fM, \tag{7.25}$$

which is to be minimized subject to the constraints

$$\int_X u[q(x, y)] m(y) dy + z(x) = U^* \quad \text{for all } x \text{ such that } n(x) > 0 \tag{7.26}$$

$$m(x) + n(x) \leq 1 \quad x \in X \tag{7.27}$$

$$\int_X n(x) dx = N \tag{7.28}$$

$$\int_X m(y) dy = M \tag{7.29}$$

plus the standard nonnegativity constraints. This is equivalent to maximizing $S = NY - C$ subject to the same constraints. Solving (7.26) with respect to $z(x)$, substituting the solution in (7.25), and using (7.28) and (7.29), we may rewrite S as follows:

$$S = \int_X \left\{ \int_X [u[q(x, y)] - q(x, y)\,(c + t|x - y|)]\, m(y) dy \right\} n(x) dx \\ + N(Y - U^* - R_A) - (R_A + f)M \tag{7.30}$$

subject to (7.27)–(7.29).

The bracketed term in (7.30) may be maximized with respect to each $q(x, y)$. Under (7.4), this yields

$$q^o(x, y) = \alpha\beta \exp -\alpha[c + t|x - y|] \quad x, y \in X. \tag{7.31}$$

Comparing (7.5) and (7.31), as expected, we find that the optimal consumption is given by the equilibrium consumption when $p(y) = c$. Substituting (7.31) into (7.30), we obtain

$$S = \gamma_1 \int_X \int_X m(y)n(x) \exp(-\alpha t|x - y|)\, dy dx$$
$$+ N(Y - U^* - R_A) - (R_A + f)M, \tag{7.32}$$

where $\gamma_1 \equiv \beta \exp -\alpha c$. The double integral stands for the sum across consumers of the indirect utility derived from the differentiated product priced at marginal cost. Because $N(Y - U^* - R_A) - (R_A + f)M$ is a constant, our problem amounts to maximizing

$$\gamma \int_X \int_X m(y)n(x) \exp(-\alpha t|x - y|)\, dy dx$$

with respect to $m(\cdot)$ and $n(\cdot)$ subject to (7.27)–(7.29). In this expression, we have replaced γ_1 by $\gamma \equiv \gamma_1/e$ for reasons that will become clear just below.

Applying the maximum principle of optimal control theory shows that a multiplier function $R^o(x)$ associated with (7.27) and two multipliers U^o and π^o associated with (7.28) and (7.29), respectively, exist such that the following conditions hold for the optimal densities $m^o(\cdot)$ and $n^o(\cdot)$:

$$R^o(x) = \max\left\{\gamma \int_X m^o(y) \exp(-\alpha t|x - y|)\, dy - U^o,\right.$$
$$\left.\gamma \int_X n^o(x) \exp(-\alpha t|x - y|))\, dx - \pi^o, R_A\right\} \tag{7.33}$$

$$\gamma \int_X n^o(y) \exp(-\alpha t|x - y|)\, dy - \pi^o = R^o(x) \quad \text{if } n^o(x) > 0 \tag{7.34}$$

$$\gamma \int_X m^o(y) \exp(-\alpha t|x - y|)\, dy - U^o = R^o(x) \quad \text{if } m^o(x) > 0 \tag{7.35}$$

$$m^o(x) + n^o(x) = 1 \quad \text{if } R^o(x) > R_A \tag{7.36}$$

in addition to (7.27)–(7.29).

Intuitively, these conditions can be explained as follows. In our setting, there are three activities: consumption, production, and agriculture. If we marginally increase the number of consumers at x, this leads to an increase in the objective function by an amount given by $\gamma \int_X m^o(y) \exp(-\alpha t|x - y|) dy$. However, this incremental benefit must be reduced by the value of the multiplier associated with the consumer population constraint. The same argument holds, mutatis mutandis, if we marginally increase the number of firms at x. Because one consumer or one firm uses one unit of land (7.33) means that $R^o(x)$ is equal to the highest marginal value of land at x. Condition (7.34) means that consumers are located at x provided that their marginal value of land is the highest one; the same applies to (7.35). Finally, (7.36) says that all the land at x is used by firms, households, or both when their marginal value of land exceeds R_A.

Clearly, we do not change the optimal densities $n^o(x)$ and $m^o(x)$ when we replace U^o with $U^o - Y$ and π^o with $\pi^o + f$ in (7.33)–(7.35). Then, we see that the optimality conditions (7.33)–(7.36) are identical to the equilibrium conditions (7.11)–(7.14) when U^o is replaced by U^* and π^o by π^*. Consequently, the equilibrium land use is identical to the optimum land use.

This is a rather surprising result because firms price above marginal cost. That the market leads to a lower consumption of the M varieties does not prevent the market from yielding the optimal pattern of land use. Of course, we must stress the following three assumptions: (1) the number of firms is fixed and the same in both the equilibrium and the optimum, (2) there is no substitution between land and consumption goods, and (3) there is no substitution between land and capital (Liu and Fujita 1991).

It is even more surprising that the equilibrium land use pattern is strongly inferior to the optimal one when firms are constrained to price at marginal cost. Indeed, because operating profits are now zero at all locations, the firms' bid rent curve is flat across locations and equal to R_A because firms minimize their loss (recall that firms are forced to be active). In this case, the equilibrium pattern involves a residential area integrated with some firms surrounded by two business sections (even when $M < N$). Regardless of the values of N and M, the density of firms in the integrated area is always much lower than at the optimum. Accordingly, pricing at marginal cost results in a much more dispersed configuration of firms.[8]

That the land use pattern is not optimal when firms price at marginal cost does not contradict the first theorem of welfare economics. Indeed, although now all prices are given to the agents, our equilibrium concept is not competitive. To find his utility-maximizing location, each consumer must know what the entire distribution of shopping opportunities will be (see (7.7)). Similarly, to find its profit-maximizing location, each firm must know what will be its aggregate demand, which depends itself on the whole consumer distribution over space (see (7.8)). However, there is no price accounting for the difference in the accessibility of one agent to the rest of the economy.[9] Therefore, the information needed by the agents to be able to *always* choose their optimal location goes beyond the usual type of information conveyed by the price system. What is here required on each agent's part has a game-theoretic flavor that brings us far away from the competitive paradigm. This shows, once more, that a spatial economy cannot be completely described by a system of competitive markets.

This negative conclusion has powerful implications, however. It is the discrepancy between price and marginal cost that allows firms to compete with other agents on the land market and to sustain the agglomeration of firms. When price equals marginal cost, the firms' incentive to locate close to consumers vanishes. Not surprisingly, therefore, the market outcome is too dispersed because there is no longer any agglomeration force.

7.2.3 Free Entry

In the previous section, the mass M of firms (or varieties) has been considered as given. The free-entry condition implies that the equilibrium profit level $\pi^* = 0$. This condition in turn allows for the determination of the equilibrium number M^* of firms. The optimum problem studied in the preceding section may easily be extended to the determination of the optimal number M^o of firms. As is well known from various models in industrial organization (see, e.g., Spence 1976, Dixit and Stiglitz 1977, Lancaster 1979, and Salop 1979), the equilibrium and optimum numbers of firms often differ. It is not even clear that one number is almost always larger than the other. In the present context, it can be shown that $M^o > M^*$, as in Dixit and Stiglitz (1977).[10] Because the model used here is similar to theirs, this result should not come as a surprise.

7.3 OLIGOPOLISTIC COMPETITION AND THE AGGLOMERATION OF RETAILERS

7.3.1 Spatial Competition under Preference for Variety

Until now, we have assumed a continuum of firms, thus implying that there is no strategic interaction on the product market. This turns out to be a convenient framework to study the working of the product market in relation to other markets such as the land or labor market or both. However, the monopolistic competition model fails to account for the strategic aspects that spatial proximity brings about (see Section 4.5).

There is an old tradition in location theory, going back at least to Hotelling (1929), which suggests that spatial competition leads to the agglomeration of firms. In the typical example of two vendors selling a homogeneous product, each firm gains by establishing near its competitor on the more populated side of the market. The only equilibrium is then obtained when both firms are located at the median of the consumer distribution where no additional gains are possible provided that transport costs increase with distance. In the case of a uniform density, the median becomes the market center. Hence, two firms competing for clients choose to minimize their spatial differentiation.[11]

The proponents of this approach, however, overlooked the fact that firms selling a homogeneous product always want to locate far apart to avoid the devastating effects of price war. Indeed, when (at least) two firms are located back to back, they get trapped into a Bertrand situation in which they find themselves with zero operating profits. This cannot be an equilibrium because firms could restore positive profits by moving away unilaterally and exploiting the monopoly power each firm has on the consumers situated in its close vicinity. For example, when transport costs are quadratic in distance and consumers are evenly distributed with unit density, the equilibrium prices of any second stage subgame decrease with the interfirm distance, whereas the equilibrium locations of the first-stage game are given by $y_1^* = 0$ and $y_2^* = l$ (d'Aspremont, Gabszewicz, and Thisse 1979).

This extreme spatial dispersion is the result of a trade-off in which price competition pushes firms away from each other whereas competition for market area tends to pull them together. To illustrate how this trade-off works, let π_1^* be firm 1's profit evaluated at the equilibrium prices $p_i^*(y_1, y_2)$ corresponding to the location pair $y_1 < y_2$. Then, because $\partial\pi_1/\partial p_1 = 0$, we have

$$\frac{d\pi_1^*}{dy_1} = \frac{\partial\pi_1}{\partial p_2}\frac{\partial p_2^*}{\partial y_1} + \frac{\partial\pi_1^*}{\partial y_1}.$$

In general, the terms on the right-hand side of this expression can be signed as follows. The first one corresponds to the *strategic effect* (the desire to relax price competition) and is expressed by the impact that a change in firm 1's location has on price competition. Because goods are spatially differentiated, they are substitutes and thus $\partial\pi_1/\partial p_2$ is positive; because goods become closer substitutes when y_1 increases, $\partial p_2^*/\partial y_1$ is negative. Hence, the first term is negative. The second term, which corresponds to the *market area effect*, is positive. Consequently, the impact of reducing the interfirm distance upon firms' profits is undetermined. However, when firms are close enough, the first term always dominates the second, and thus firms always want to be separated in the geographical space. This implies that the principle of minimum differentiation ceases to hold when firms are allowed to compete in prices (d'Aspremont, Gabszewicz, and Thisse 1983). Accordingly, when firms choose locations and prices sequentially, price competition is a strong dispersion force sufficient to destroy agglomeration in spatial competition.[12]

When $M \geq 2$ firms choose prices and locations simultaneously, a Nash equilibrium never exists (at least in pure strategies), that is, an M-tuple of prices and an M-tuple of locations obtain such that no firm can strictly increase its profits by unilaterally changing its location, price, or both. Consider indeed any two firms and suppose that such an equilibrium exists. Whatever the market configuration, the firm earning (weakly) smaller profits could strictly increase its profits by locating at the same place as the other firm and slightly undercutting its price. By doing this, the undercutting firm usurps the profits of its rival while still retaining a part of its previous market. It is therefore strictly better off, a contradiction to the equilibrium condition.

However, these negative results do not kill the subject. It should be kept in mind, indeed, that they are based on an extreme price sensitivity of consumers: if two firms are located side by side with identical prices, a small price reduction of one firm will attract all the customers. Such extreme behavior seems unwarranted. When the product is differentiated and when consumers like product variety, the aggregate response to a price cut will not be so abrupt because the quality of product match matters to consumers. Product differentiation then alleviates price competition. This modification of the spatial competition model, which has been developed by de Palma et al. (1985), has two major implications. First, if consumers' preference for variety becomes sufficiently large, firms' demand functions are smoothed sufficiently even when they are located close together so that a price equilibrium in pure strategies exists. Second, under the same condition, firms tend to agglomerate at the market center to have the best access to the market, as suggested by Hotelling. Price competition at the center is relaxed because of the differentiation among vendors, which gives them market power even when they are agglomerated. As we will see, agglomeration can then be shown to be a Nash equilibrium when transportation costs are low with respect to product differentiation.

Hence, we follow the main idea of the previous section by assuming that firms sell a differentiated product and that consumers like variety. Nevertheless, unlike before, we consider a finite number M of firms behaving strategically. We will also follow the Hotelling tradition by assuming that the consumer distribution over the location space $X = [0, l]$ is fixed. When each consumer uses one unit of land, the consumer distribution is uniform over X. Like Hotelling again, it is supposed that firms do not consume land. As in Section 7.2, the utility of a consumer is additive:

$$U(z; q_i, i = 1, \ldots, M) = \sum_{i=1}^{M} u(q_i) + z.$$

The main difference with (7.1) is that we have here a finite number of firms instead of a continuum. We saw in Section 7.2 that there is no direct substitution among varieties. To introduce this effect explicitly in a simple way, we suppose that each consumer buys a fixed number $\bar{q} > 0$ of units of the differentiated product per unit of time (e.g., a given number of restaurant dinners per month):[13]

$$\sum_{i=1}^{M} q_i = \bar{q}. \tag{7.37}$$

Such a constraint implies that firms compete for clients within a market of a given size. Here also, we suppose that the utility U has an entropy-form:[14]

$$U = \begin{cases} \sum_{i=1}^{M} q_i - \sum_{i=1}^{M} \frac{q_i}{\alpha} \log q_i + z & \text{if } \sum_{i=1}^{M} q_i = \bar{q} \\ -\infty & \text{otherwise} \end{cases}. \tag{7.38}$$

As usual, the budget constraint of a consumer at $x \in X$ is

$$\sum_{i=1}^{M}(p_i + t|x - y_i|)q_i + z = Y,$$

where p_i is the (mill) price selected by firm i selling variety i, and $y_i \in X$ is the location chosen by this firm.

Using standard optimization techniques, we get[15]:

$$q_i^*(x) = \frac{\exp -\alpha(p_i + t|x - y_i|)}{\sum_{j=1}^{M} \exp -\alpha(p_j + t|x - y_j|)}\bar{q} \qquad x, y_i \in X, \tag{7.39}$$

which, unlike (7.5), depends not only upon the firm's price p_i and location y_i but also upon the prices and the locations chosen by all its rivals ($j \neq i$). Indeed, (7.37) imposes that the total consumption of the differentiated product is fixed so that, because the number of firms is finite, the consumption of a variety impacts on the consumption of the others.

Unlike the monopolistic competition model in which price elasticity is location independent and equal to αp_i, price elasticity is now location dependent because it equals $\alpha p_i[1 - P_i(x)]$, where

$$P_i(x) \equiv \frac{\exp -\alpha(p_i + t|x - y_i|)}{\sum_{j=1}^{M} \exp - \alpha(p_j + t|x - y_j|)}$$

is known as the *multinomial logit* (McFadden 1974).[16] It is assumed here that each firm is aware of this fact when selecting its price and location. As a consequence, there is strategic interaction between firms in both prices and locations. As will be seen, this leads to a more involved and richer pattern of interdependence among firms.

To develop some insights about this interaction pattern, we first discuss the special case of two firms ($M = 2$) located at $0 < y_1 < y_2 < l$ and pricing at the same level p. Then, it can readily be verified that

$$\begin{aligned} q_1^*(x) &= \frac{\bar{q}}{1 + \exp - \alpha t(y_2 - y_1)} & x \in [0, y_1] \\ &= \frac{\bar{q}}{1 + \exp - \alpha t[y_2 - y_1 + 2(y_1 - x)]} & x \in [y_1, y_2] \\ &= \frac{\bar{q}}{1 + \exp \alpha t(y_2 - y_1)} & x \in [y_2, l]. \end{aligned}$$

Consider Figure 7.2 in which $q_1^*(x)$ is described as a function of the consumer location. We see that the demand of variety 1 is continuous over the entire location space X, constant in the two hinterlands $[0, y_1]$ and $[y_2, l]$, and decreasing in the contention segment $[y_1, y_2]$. Stated differently, the demand to store 1 is highest over its hinterland, decreasing as x moves away from y_1 to get closer to y_2, and lowest over its competitor's hinterland. Furthermore, it is easy

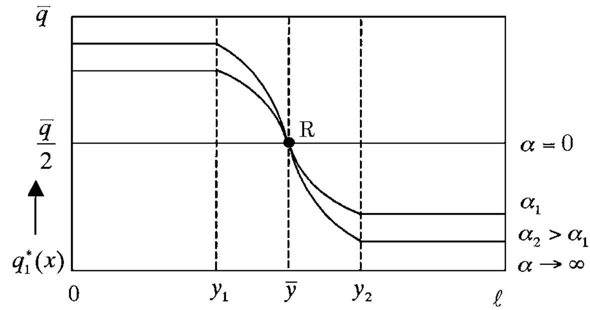


Figure 7.2: The equilibrium demand pattern to store 1 when consumers love variety.

to check that $q_1^*(x)$ is concave on $[y_1, \bar{y}]$ and convex on $[\bar{y}, y_2]$, where $\bar{y}$ is the middle point of the two suppliers. Finally, $q_1^*(x)$ exceeds $q_2^*(x)$ if and only if x is closer to store 1 than to store 2. At $x = \bar{y}$, both demands are equal to $\bar{q}/2$. The details of this demand pattern agree with intuition and experience.

When the degree of differentiation, measured by $1/\alpha$, increases, the quantity $q_1^*(x)$ shifts downward (upward) on the left (right) of $\bar{y}$. This occurs because consumers value more product diversity and are, therefore, less sensitive to spatial proximity. In the limit, when $\alpha = 0$, distance becomes inessential and $q_1^*(x) = 1/2$ for all $x \in X$. On the other hand, when $\alpha \to \infty$ (the product is homogeneous), $q_1^*(x) = 1$ for $x \in [0, \bar{y})$ and 0 for $x \in (\bar{y}, l]$, that is, each consumer patronizes the closer store; we fall back on Hotelling's original model.

Consider now Figure 7.3 in which a single firm locates at y_1 and $(M - 1)$ firms at y_2. It is straightforward to show that

$$
\begin{aligned}
q_1^*(x) &= \frac{\bar{q}}{1 + (M-1)\exp - \alpha t(y_2 - y_1)} && x \in [0, y_1] \\
&= \frac{\bar{q}}{1 + (M-1)\exp - \alpha t[y_2 - y_1 + 2(y_1 - x)]} && x \in [y_1, y_2] \\
&= \frac{\bar{q}}{1 + (M-1)\exp \alpha t(y_2 - y_1)} && x \in [y_2, l].
\end{aligned}
$$

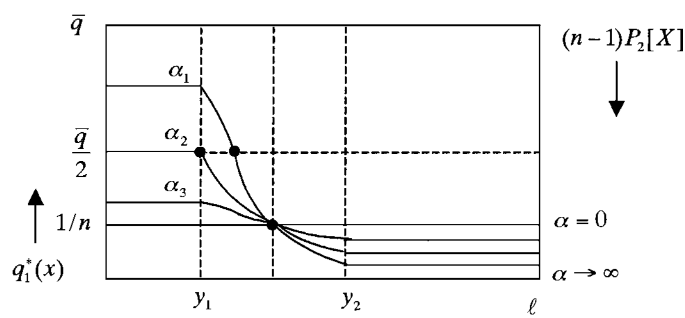


Figure 7.3: The equilibrium demand pattern to store 1 when the other stores are agglomerated at y_2.

Clearly, the demand to firm 1 falls as the number of firms agglomerated at y_2 rises, thus making the clustering of firms more attractive as a whole. When the number of firms at y_2 is not large, or when the degree of product differentiation is not high, or both, store 1 has a demand over its hinterland and some part of the contention segment, which exceeds the one at the market center (where the other firms are agglomerated). However, when there are enough firms at y_2, or varieties are differentiated enough, or both, this advantage of the isolated firm tends to vanish, and the market center becomes increasingly attractive.

It is worth noting that this behavior agrees with the *gravity principle* developed in retailing models by Reilly (1931) according to which distance between places is an impediment to interaction whereas the size of a place makes it more attractive to consumers. This principle has been extended to the description of actual trip patterns by geographers and regional scientists and has generated a vast and rich literature known as *spatial interaction theory*.[17]

As usual, firm i's profits are

$$\pi_i(\boldsymbol{p}; \boldsymbol{y}) = (p_i - c) \int_0^l q_i^*(x) dx,$$

where $q_i^*(x)$ is given by (7.39). Assume for the moment that all prices p_i are equal and fixed. Clearly, maximizing profits amounts to maximizing demand. It can be shown that the agglomeration of M firms at $l/2$ is a Nash equilibrium if and only if[18]

$$1/\alpha t l \geq (1 - 2/M)/2. \tag{7.40}$$

Hence, for given values of α and t, increasing the number of firms makes the tendency toward central agglomeration weaker. Indeed, when M rises, the elasticity decreases at each point (since $P_i(x)$ decreases at all x) and the difference between elasticity at the agglomeration and elasticity at a noncentral location increases. As a result, the benefits of exploiting a local market may well exceed those associated with a central location.

However, such an effect can be offset by a rise in product differentiation, a fall in transportation cost, or both. Because the right-hand side of (7.40) is bounded above by $1/2$, it is apparent that the central agglomeration is always an equilibrium for any large number of stores as long as $\alpha t l \leq 2$. This is reminiscent of the results obtained in Section 6.3.3 for "small cities" provided that the degree of product differentiation plays the role of the intensity of communication between firms.

Clearly, for agglomeration to arise, varieties must be differentiated enough when $M > 2$. Consider, indeed, the extreme case in which all varieties are perfect substitutes. Then, if $M > 2$, any clustering of firms at the market center (or somewhere else) is not an equilibrium. This occurs because any firm can always substantially increase its sales by locating slightly away from the cluster (on the larger side of the market if the cluster is not established at the market center). Because the product is homogeneous, all consumers closer to the deviating firm than to the cluster will purchase from this firm so that this one can guarantee to itself a market share equal to $1/2 - \varepsilon$ instead of getting $1/M$ in the cluster. Nevertheless, the benefits of such a deviation decrease as varieties become increasingly differentiated because consumers ill-matched to the variety supplied by the deviating firm will find it advantageous to go their way as far as the cluster.

For the more general case in which firms compete in both prices and locations, we have the following:

Proposition 7.2 *Consider M firms competing in mill prices and locations. Then,* $p_i^* = c + M/\alpha(M-1)$ *and* $y_i^* = l/2$, $i = 1, \ldots, M$, *constitute a Nash equilibrium when* $\alpha tl \leq 2$.[19]

The structure of the proof is as follows (details are given in Part B of the chapter appendix).[20] When all firms are agglomerated, it is easy to see that there is a unique price equilibrium at which the common price is $p^* = c + M/\alpha(M-1)$. Assume now that $M - 1$ firms are set up at the market center and charge p^*, whereas firm 1 is located at $y_1 < l/2$, and set $p_1 \geq c$. Whatever the value of p_1, if $\alpha tl \leq 2$, firm 1's profit is increasing in y_1 over $[0, l/2)$, and so firm 1 wants to join the others at the market center. Because p^* is the only price equilibrium when all firms are together, firm 1's profit is greatest when $p_1 = p^*$ and $y_1 = l/2$. As in the preceding section, we therefore see that a high degree of product differentiation, a low transportation rate or both, sustain the agglomeration of the M stores at the market center. It can also be shown that the market center is the only agglomerated location equilibrium.[21]

When transport costs are low, the benefits of geographical separation are reduced and prices are lower.[22] Firms might then choose to reconstruct their profit margins by differentiating their products in terms of some nongeographical characteristics that are tangible or intangible. Stated differently, product differentiation is substituted for geographical dispersion (this is shown in a model of spatial competition by Irmen and Thisse 1998). In this case, firms no longer fear the effects of price competition (the centrifugal force is weakened by product differentiation) and strive to be as close as possible to the consumers with whom the matching is the best. Because these consumers are spread all over the market space, firms set up at the market center and, therefore, minimize their geographical differentiation. This result agrees with market potential theory, as developed by Harris (1954) in classical economic geography according to which firms tend to locate where they have the "best" access to markets in which they can sell their product.

Consider now the implications of the logit model (7.39) for the sequential Hotelling duopoly model. Anderson et al. (1992, chap. 9) have shown the existence and the uniqueness of a price equilibrium for any location pair when αtl is sufficiently small. Using this price equilibrium, these authors were then able to study the location game by appealing to numerical analysis. The following

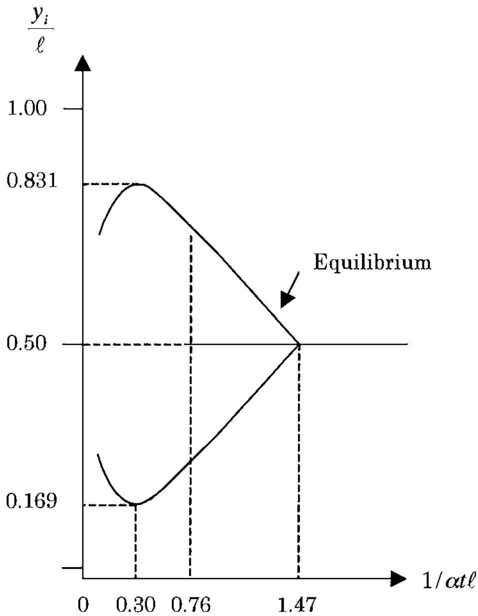


Figure 7.4: Firms' equilibrium locations when consumers love variety.

results emerge (they are depicted in Figure 7.4). As $1/\alpha tl$ rises from 0 to 0.30, there is no location equilibrium (in pure strategies). For $0.30 < 1/\alpha tl < 1.47$, there is a symmetric dispersed equilibrium that initially entails increasing geographical separation of firms. However, when $1/\alpha tl$ goes beyond some threshold (around 0.76), the geographical separation starts to decrease. For $0.76 < /\alpha tl < 1.47$, an agglomerated equilibrium exists along with the dispersed one; however, the former is unstable whereas the latter is stable. Finally, for $1/\alpha tl \geq 1.47$, there is a unique equilibrium that involves central agglomeration.

The intuition behind these results is pretty straightforward. An arbitrarily small amount of differentiation is not sufficient to restore existence because consumers' shopping behavior, though smooth, remains very sharp (i.e., close to the standard $0 - 1$ behavior). When existence is guaranteed, firms' market areas overlap, thus making price competition so fierce that firms want to move apart. Beyond some threshold, the product differentiation effect tends to dominate the price competition effect, and firms set up closer to the market center because price competition is relaxed. Finally, for a sufficiently large degree of differentiation, the market area effect becomes predominant and the agglomeration of sellers is the market outcome, as in the nonprice competition context. In both the simultaneous and sequential games, the message is the same: agglomeration arises when price competition is relaxed through sufficient product differentiation.[23]

7.3.2 The Efficiency of the Agglomeration of Firms

The welfare analysis reveals some unexpected results. Because consumers' locations are fixed, the utility level across locations may not be the same, and we cannot minimize total costs anymore. In optimum, prices are set equal to the common marginal cost c so that consumers' well-being depends only upon the firms' locations ($\boldsymbol{y}$). In the homogeneous product case, total costs are simply given by aggregate transportation costs. However, once we introduce differentiation across varieties, consumers no longer patronize the nearest firm on each trip (recall that all prices are equal to c) because they now benefit from intrinsic differentiation between shops. In this context, one needs a more general approach accounting for both distance and product diversity effects. The appropriate measure is the indirect utility. For a consumer at x, it is obtained by introducing the quantities (7.39) in the utility (7.38):

$$V(x;\boldsymbol{y}) = \frac{\bar{q}}{\alpha} \ln \left[\sum_{i=1}^{M} \exp - \alpha(c + t|x - y_i|) \right].$$

Consumer surplus is defined by the sum of the individual indirect utilities:

$$S(\boldsymbol{y}) = \int_X V(x;\boldsymbol{y})dx \tag{7.41}$$

because the density is assumed to be uniform.

Somewhat surprisingly, the optimum involves the agglomeration of firms as long as $\alpha t l \leq 2$. In what follows, we give a proof for the case in which $M = 2$. It is intuitively plausible (and it can be shown) that the pair of optimal locations is symmetric. Hence, we may set $y_1 = l/2 - a$ and $y_2 = l/2 + a$ and rewrite the consumer surplus (7.41) as follows:

$$\begin{aligned} S(a) = {} & \frac{2}{\alpha} \int_{l/2}^{l/2+a} \ln[\exp - \alpha t(x - l/2 + a) + \exp -\alpha t(l/2 + a - x)]\, dx \\ & + \frac{2}{\alpha} \int_{l/2+a}^{l} \ln[\exp - \alpha t(x - l/2 + a) + \exp -\alpha t(x - l/2 - a)]\, dx. \end{aligned}$$

After some routine manipulations, we obtain

$$\begin{aligned} S(a) = {} & -t[l/4 - a(1 - 2a)] + \frac{2}{\alpha} \int_{l/2+a}^{l} \ln[1 + \exp -2\alpha t(x - l/2)]\, dx \\ & + \frac{2}{\alpha}(l/2 - a) \ln[1 + \exp 2\alpha t a]. \end{aligned} \tag{7.42}$$

Differentiating (7.42) with respect to a leads to

$$1 - 4a - \exp(-2\alpha t a) = 0. \tag{7.43}$$

Clearly $a = 0$ is always a root of this equation. A necessary and sufficient condition for (7.42) to have a strictly positive root is that the derivative of $\exp(-2\alpha t a)$ must be smaller than the derivative of $1 - 4a$, both evaluated at $a = 0$. This is equivalent to $\alpha t l > 2$. Differentiating (7.43) with respect to a and using the second-order condition, we get

$$\frac{da^{\circ}}{d\alpha} > 0.$$

Hence, as the degree of product differentiation rises (α decreases), we see that the optimal distance between the two stores falls. Because distance matters less and less to consumers relative to preference for variety, moving the stores toward more central locations becomes increasingly desirable; this allows one to increase the accessibility of consumers dispersed over the entire segment to more variety.

On the other hand, when $\alpha t l \leq 2$, the first derivative of $S(a)$ is negative for all admissible values of a, thus implying that $S(a)$ is maximized at $a = 0$. In other words, when the varieties are differentiated enough, the two stores must be located at the market center to maximize the consumer surplus. This occurs because the store's attributes dominate the transportation factor sufficiently to render the market center desirable from the consumer's standpoint (see de Palma, Liu, and Thisse 1994 for more details).

It remains to deal with profits. As usual, we assume free entry. Anderson et al. (1992, chap. 7) have shown that the optimal number of firms is equal to the equilibrium number minus 1. Therefore, we may conclude that the formation and size of the cluster of stores at the market center are (almost) socially optimal when product differentiation is strong enough ($\alpha t l \leq 2$).

As noticed by Eaton and Lipsey (1979, 21), "it is part of conventional wisdom of economists to refer to Hotelling and to the socially wasteful nature of clustering whenever local clusters are encountered in the real world." Our results do not confirm this view because we have just seen that the formation of clusters is socially desirable when varieties are differentiated enough, when transport costs are low enough, or both, that is, $\alpha t l \leq 2$. The conventional wisdom is wrong here because it ignores the fact that firms sell differentiated products whereas consumers have heterogeneous tastes. However, when $\alpha t l > 2$, it is not optimal to group firms anymore, and thus high transportation costs result in separate locations. In this case, Anderson et al. (1992, chap. 9) have provided several examples suggesting that the market may well provide insufficient geographical dispersion.

Note, finally, the following relationship with the model of Section 7.2. When firms' land consumption (S_{f}) shrinks to zero, there is only one equilibrium involving the agglomeration of all firms at the city center. This opens the door to the possibility of extending the present model by adding a land market (as in Section 4.5.3). Furthermore, the common equilibrium price (given in

Proposition 7.2) is identical to (7.6) when $M \to \infty$, suggesting that the monopolistic competition model of Section 7.2 can be viewed as the asymptotic version of the spatial competition model discussed here.

7.3.3 Shipping Goods and Agglomeration

Shipping models come from a different tradition that derives from the analysis of spatial price discrimination in an oligopolistic environment (local markets are segmented, whereas they are tied in shopping models). When products are delivered, the location of the customers is observable to the firms, which are then able to price discriminate across locations (this corresponds to third-degree price discrimination à la Pigou). Shipping models were initiated by Hoover (1937) and have been much developed in the last two decades.[24]

Consider the case in which firms compete in price schedules. This means that each firm announces for each location a *delivered price* at which it is willing to supply the corresponding customers (e.g., pizzerias). Then, discriminating firms always want to be located far apart when the product sold is homogeneous (Lederer and Hurter 1986). This is again because price competition at each consumer location is such a strong centrifugal force that firms are hurt by geographical proximity. However, for exactly the same reasons as those discussed in Section 7.3.1, discriminating oligopolists locate closer when they supply a product that is more differentiated (Anderson and de Palma 1988).

This tendency toward agglomeration is even stronger when firms compete in quantity schedules. Indeed, as shown by Anderson and Neven (1991), when the product is homogeneous, agglomeration at the market center is the unique equilibrium for any given number M of firms if $tl \leq (1 - c)/M$, that is, if shipping costs are sufficiently low. Hence, although the product is homogeneous, firms do agglomerate because quantity-setting firms are less affected by competition, making the market area effect dominant.

However, when t goes beyond some threshold, the center is no longer an equilibrium: firms want to differentiate their locations in order to retain enough customers near the market endpoints. This new effect, called the market periphery effect, corresponds to a centrifugal force. It is present in any spatial competition model in which local demands are price sensitive. When transportation costs are low, this force is not strong enough to prevent the market area effect from dominating, thus leading to agglomeration. The opposite holds when transportation costs get higher: the market equilibrium involves gradual dispersion of producers. As in the shopping models both agglomerated and dispersed equilibria may coexist for some range of t (Gupta, Pal, and Sarkar 1997).

Accordingly, although shopping and shipping models have different aims and obey different incentive systems, it seems fair to say that they are governed by essentially the same centrifugal and centripetal forces, thus leading to similar locational patterns under similar conditions.

7.4 CONSUMERS' SEARCH AND THE CLUSTERING OF SHOPS

As in Section 7.3, the population of consumers is uniformly distributed along the segment $[0, l]$, whereas firms sell differentiated varieties of the same product. However, although the typical consumer knows which varieties are available in the market, he is unsure about which variety is offered where and at which price.[25] Because consumers must compare alternatives before buying, they must undertake search among firms. More precisely, it is assumed that the only way for them to find out which variety is on offer in a particular store is to visit this store and to pay the corresponding transportation cost. Indeed, it is hard to figure out how good the matching with a product is without seeing it. Collecting information by telephone may be helpful for prices but not for varieties. Given the expense of gathering information, each consumer must compare the cost of an additional bit of information with the expected gain in terms of expected surplus. In a spatial setting, both vary with the consumers' and firms' locations.[26]

In such a context, when several firms are located together, it is reasonable to assume that the typical consumer knows the location and the size of the cluster but not its composition. Once the consumer arrives at the cluster, the travel costs are sunk, and he can visit any store at a very low cost. In other words, each consumer visiting the cluster enjoys scope economies in search. On the other hand, the consumer must pay the transport cost to each isolated store he visits. Geographical clustering of stores is therefore a particular means by which firms can facilitate consumer search. Indeed, a consumer is more likely to visit a cluster of stores than an isolated one because of the higher probability he faces of finding a good match and a good price there. When firms realize this fact, each of them understands that it might be in its own interest to form a marketplace with others.

Matters are not that simple, however. As observed by Stahl (1982, 98),

> the aggregate demand observed at a marketplace increases with its size, as defined by the number of commodities offered there. Therefore, a seller, upon choosing a profit maximizing location, is confronted with the alternative either to establish a local monopoly with a small market area and a large share of consumers purchasing, by lack of variety, his variety; or to join other firms in a competitive marketplace with a large market area in which he fetches the demand from only the small subset of consumers not substituting away towards the other alternatives available there.

In other words, when a firm considers the possibility of joining competitors within the same marketplace, it faces a trade-off between the following two effects: a negative competition effect and a positive market area effect, both being generated by the pooling of firms selling similar products. As we will see in Chapter 9, a similar effect is at work in several models developed in economic geography.

In this section, we follow the analysis of Wolinsky (1983) and show how a cluster of firms may emerge as a Nash equilibrium when firms sell their varieties at a common fixed price. Product differentiation is modeled using a spatial setting à la Hotelling–Lancaster. More precisely, varieties of the same product are evenly distributed along a circle C of unit length, thus substituting a Lancasterian space of characteristics to the geographical space used in Section 4.5 (see, e.g., Lancaster 1979 and Salop 1979). The location r_i of firm i now stands for the position of its variety in the characteristic space C, the location r of a consumer for his ideal product, whereas the transportation cost $s|r - r_i|$ corresponds to the utility loss incurred for not consuming his ideal product, where $|r - r_i|$ stands for the length of the shorter arc between r and r_i. As a result, a consumer is now described by two parameters: his location x in the geographical space $[0, l]$ and his ideal product r in the characteristics space C. The two distributions are supposed to be independent so that the distribution of consumer type (x, r) is uniform over the cylinder whose basis is C and height l.

Nonconvexities in transportation are needed for the cluster to emerge here. The simplest form is to assume that consumers bear some positive fixed cost t_0 each time they make a separate trip because of the corresponding terminal conditions (e.g., parking, waiting time for a bus). This is certainly a reasonable assumption and, if we did not make it earlier it is because the existence of a positive fixed cost in transportation has no impact on the results obtained so far inasmuch as consumers visit a *single* place each time they make a purchase. On the other hand, as will be seen in Proposition 7.3, this turns out be to be critical for the description of a consumer's search strategy. Furthermore, instead of considering $t|x - y|$ as the cost of a round trip between x and y, it now describes a one-way trip. This assumption is made because the return trip may differ from the initial trip. Hence, if consumers were to know which firm sells which variety, the indirect utility of a consumer of type (x, r) patronizing firm i would be given by

$$V_i(r, x) = Y - p + u - s|r - r_i| - 2\,(t_0 + t|x - y_i|),$$

where p is the common fixed price, s the marginal utility loss for not consuming at one's ideal product (also called the matching cost), y_i firm i's endogenous location, and t_0 the fixed transport cost. However, although consumers are able to observe the location y_i of firm i, they do not observe the variety (r_i) it sells. For that, consumers must visit firm i and bear the corresponding travel cost.

Because we want a cluster formed by all firms to be an equilibrium, it is sufficient to investigate the case in which $M - 1$ firms are located together at y_C, whereas the remaining firm (say firm M) is alone at $y_I \neq y_C$, and to show that this firm is better off by joining the cluster. For simplicity in notation, we set $\Delta \equiv |y_C - y_I| > 0$. In the case of such a configuration, the search plan of a consumer then consists of a decision made on the basis of two things: (1) where to start and (2) when to stop the search.

A consumer has two possible plans: either he starts at the cluster and, possibly, continues to the isolated firm, or proceeds the other way round. In both cases, the consumer will adopt a sequential search with a fixed stopping rule. Because prices are fixed and identical across stores, the only element that matters in the consumers' decision whether to continue a search is the quality of the match between the searched varieties and the ideal product. The optimal search is therefore to keep on searching until a variety within the "reservation distance" in the characteristics space C is found. This means that the consumer buys from the *first* store offering the variety whose distance to the ideal product is less than, or equal to, the distance D at which the expected utility increase from sampling another store is just equal to the additional search cost (see below for the formal definition of the reservation distance). When the consumer visits a shop in which the variety does not fall within his reservation distance, the search is continued (McMillan and Rothschild 1994).

Suppose that the consumer (x, r) first visits the cluster. When visiting a new store within the cluster, this consumer must bear a cost k independent of his tastes and location. Because all consumers know that varieties are equidistantly located along the circle C, consumers behave approximately as if the distribution of varieties were uniform along C when the number M of varieties is large while evaluating the benefit of a further search.

When the consumer has already visited some stores within the cluster and the best-offered variety is at distance D from the ideal product, the expected utility gain from visiting another store in the cluster is defined by the following expression (recall that varieties are supposed to be uniformly distributed along C):

$$B(D) = \int_0^D [(Y - p + u - s\delta) - (Y - p + u - sD)]\, d\delta$$
$$= \int_0^D (sD - s\delta)\, d\delta.$$

The reservation distance D_C inside the cluster is then obtained by equalizing the expected utility gain and the additional search cost, that is, D_C is the unique solution to the equation:

$$B(D) = k,$$

where k is the cost of sampling a new store in the cluster. The search is stopped when a variety at a distance smaller than, or equal to D_C is found in a store established in the cluster. In other words, the typical consumer has an "acceptance zone" centered at his or her ideal point whose size is $2D_C$. The reservation distance D_C increases with the search cost k and decreases with the matching cost s. Note that s and k must be such that $D_C = (2k/s)^{1/2} < 1/2$, for otherwise no search would be undertaken. Furthermore, a consumer buys from any store in the cluster he is going to visit with a probability equal to $1/2D_{\rm C}$ because this is the probability that the variety supplied by this store falls in the consumer's acceptance zone. Because consumers have identical preferences (up to a rotation of their ideal product) and the same beliefs about the distribution of varieties, the value of $D_{\rm C}$ is the same across consumers.

Assuming that the isolated store carries the variety at distance $D_{\rm I}$ from his ideal product, a consumer first visiting this store will buy from it if and only if the expected gain from buying in the cluster does not exceed the transportation cost to go there. The expected gain associated with a continued search at the cluster is

$$(Y - p + u - sD_{\rm C}) - [s(D_{\rm I} - D_{\rm C}) + k](1 - 1/2D_{\rm C})^{M-1} \\ - (Y - p + u - sD_{\rm I}),$$

where $1 - 1/2D_{\rm C}$ is the probability that the consumer does not find a variety within his acceptance zone in a store sampled at cost k within the cluster. Hence, the reservation distance $D_{\rm I}$ at the isolated firm is given by the solution of

$$s(D_{\rm I} - D_{\rm C}) - [s(D_{\rm I} - D_{\rm C}) + k](1 - 1/2D_{\rm C})^{M-1} = t_0 + t\Delta,$$

where $t_0 + t\Delta$ is the transport cost to the cluster. Note that a positive solution may not exist. If it does, then $D_{\rm I}$ increases with the transport cost (t_0 and t) as well as with the distance Δ between the isolated firm and the cluster. More important, it decreases with the matching cost s and the number of stores established in the cluster. As a consequence, the attractiveness of a cluster increases with its size as well as with consumers' matching cost.

The expected match is always higher when the cluster is visited first because more varieties are available there. However, the expected transport costs to the cluster and to the isolated firm, denoted respectively by $T_{\rm C}$ and $T_{\rm I}$, generally vary with the consumer's location x. The following result identifies a sufficient (but not necessary) condition on M and Δ for the difference in transport costs ($T_{\rm I} - T_{\rm C}$) to be positive for *all* consumers when the distance between the cluster and the isolated firm does not exceed $\Delta(M)$. Accordingly, when $\Delta \leq \Delta(M)$, there is no spatial search (the proof is given in Part C of the chapter appendix).

Proposition 7.3 *Consider $M - 1$ firms located at $y_{\rm C}$ and one firm at $y_{\rm I}$. If M is large enough, a distance $\Delta(M)$ exists such that all consumers first visit the cluster if the distance between the cluster and the isolated firm does not exceed $\Delta(M)$. Furthermore, $\Delta(M)$ is increasing in M.*

The rest of the argument is straightforward. Let $M > \tilde{M} \equiv \max\{\Delta(M), 1/2D_{\rm C}\}$. Then, if one firm is located at most $\Delta(M)$ away from the cluster, all consumers choose to go to the cluster first. In addition, because $M > 1/2D_{\rm C}$, each consumer finds in the cluster a variety below his reservation distance $D_{\rm C}$. This implies that the isolated firm has no customers. As a result, this firm would be strictly better off by joining the cluster where it enjoys a market share equal to $1/M$. If the urban area is small enough for $\Delta(M) \geq l/2$ to hold, all consumers therefore buy from the cluster. Accordingly, we have the following result.

Proposition 7.4 *Consider M firms selling differentiated varieties and a continuum of consumers who do not know which firm offers which variety. If* $M > \tilde{M}$ *and* $\Delta(M) \geq l/2$, *then* $y_i^* = y^* \in [\max\{0, l - \Delta(M)\}, \min\{\Delta(M) - l/2, l\}]$, *for* $i = 1, \ldots, M$, *is a Nash equilibrium.*

In other words, the ignorance of consumers about the available varieties leads to the emergence of a cluster when the size of the urban area is small (or, equivalently, when variable transport costs as measured by t are low) and when there are enough stores to make the cluster attractive to all consumers.

By now, the role played by the fixed transport cost t_0 should be clear. If this cost were equal to zero, all consumers located to the left of the isolated firm would always visit this firm before the cluster because this firm could be sampled at no cost. As a result, any single firm would have an incentive to locate close to the cluster on the larger side of the market inasmuch as it would be visited by a majority of consumers. That the emergence of a cluster is more likely when the number of firms increases agrees with the gravity principle developed in economic geography.

It is worth noting that *the agglomeration may arise away from the market center*. Indeed, Proposition 7.4 does not require the cluster to be at the point minimizing total transportation costs. Any point such that no single firm is able to find an alternative location far enough to induce some consumers to visit it before the cluster is an equilibrium even when most of the population is concentrated away from it (note that the proof of Proposition 7.3 does not use the assumption that the population is uniformly distributed along the interval $[0, l]$). To illustrate the implications of this result, consider a cluster established in the middle of the urban area while this area starts expanding leftward. Then, the cluster still attracts new firms entering the market even though more consumers are now located to the left of the cluster. By its mere existence, a cluster generates a lock-in effect similar to those that we will encounter in some subsequent chapters.

Of course, the cluster tends to be not too far from the market center because stores need to offer good accessibility to all consumers. Once the urban area extends far away into the same direction, some firms will want to create a new cluster to the left of the original one, thus leading to a (hierarchical) spatial structure of stores within the expanding urban area.

The foregoing findings can be extended to the more general case in which firms also choose prices strategically and both prices and variants are not directly observable by consumers. Wolinsky (1983) has then shown the existence of a symmetrical price equilibrium.

Schulz and Stahl (1989; 1996) have shown that it is possible to uncover additional and surprising results by considering a market of variable size. To this end, they consider an unbounded geographical space that allows them to capture the idea that *more competition within the cluster may attract more customers coming from more distant locations, thus allowing the demand for each variety to increase.* In other words, the entry of a new variety may lead to an increase in the cluster's demand that outweighs the decrease in market share inflicted on existing varieties. Furthermore, prices increase with the number of firms so that individual profits first rise and then fall with the number of firms in the cluster. Clearly, when the number of varieties is not too large, such positive effects associated with the gathering of firms strengthen the agglomeration force that lies behind the cluster.

Though several firms may collectively want to form a new market, it may not pay an individual firm to open a new market in the absence of a coordinating device. Consequently, a new firm entering the market will choose instead to join the incumbents, thus leading to a larger agglomeration. In such a setting, the entry of a new firm creates a positive externality for the existing firms by making total demand larger. Although price competition becomes fiercer, it appears here that firms take advantage of the extensive margin effect to increase their prices in equilibrium. In other words, the market size effect "transforms" goods that are substitutes in the consumers' eyes into complements competing in the same market. As observed by Eaton and Lipsey (1977), this might explain the common fact that department stores encourage the location of competing firms within the shopping center.

A related idea is explored by Gehrig (1998) when two differentiated markets are separated. Unlike Schulz and Stahl, Gehrig supposes that the aggregate demand over the two local markets is fixed. The number of products available in a local market increases with the number of consumers visiting this place, thus reducing the average matching costs. The attractiveness of market therefore depends on the size of its clientele. Gehrig then shows that, in such a setting, an entrant is likely to join one of the existing markets – especially when transportation costs are low.

Although the context differs from those considered in Chapters 4 and 5, the foregoing discussion illustrates once more the role that a land developer or a public authority, internalizing here the demand externality, may play in the emergence of a new commercial area such as a shopping mall. Stated differently, coordination failure (or a missing agent) strengthens the agglomeration force generated by the lack of information on the consumer side.[27]

7.5 THE FORMATION OF URBAN EMPLOYMENT CENTERS

7.5.1 Monopolistic Competition on Markets for Intermediate Goods

In Section 7.2, we focused on monopolistic competition on the product market. Here we want to show how the same principles can be applied to study how the availability of intermediate goods may affect the agglomeration of firms. In Section 4.3, we have analyzed the role of variety in intermediate goods under the assumption that all producers of such goods locate together in the CBD. Our purpose is to make this center endogenous by following an approach that is essentially identical to the one taken in Section 7.2 in that we replace consumers by exporting firms and firms producing consumption goods by specialized firms producing intermediate goods.

We consider again a linear space $X = (-\infty, \infty)$. There is a continuum M_e of identical firms selling a final good on the world market at a given price normalized to 1, and using a continuum M_s of specialized intermediate goods, land, and labor. These firms can be interpreted as the headquarters of multinational firms (e.g., New York), or as high-tech firms (e.g., the Silicon Valley), or as manufacturing firms (e.g., old Pittsburgh or Detroit). The production function of a firm belonging to the exporting sector is given by

$$X_e = \int_0^{M_s} v[q(i)]di \tag{7.44}$$

in addition to the use of some fixed requirements of land S_e and of labor L_e. In (7.44), $q(i)$ is the quantity of the intermediate good i, and $v(\cdot)$ expresses the contribution of any intermediate good to the output of the final sector. As in Section 7.2, we assume that v is given by an entropy-type such as

$$v(q) = \begin{cases} \frac{q}{\alpha}(1 + \log\beta) - \frac{q}{\alpha}\log\frac{q}{\alpha} & \text{if } q < \alpha\beta \\ \beta & \text{if } q \geq \alpha\beta \end{cases}.$$

Thus, the production function (7.44) exhibits increasing returns in the number of intermediate goods (the production counterpart of preference for variety in consumption). Assume, indeed, that all intermediate goods are sold at the same price $\bar{p}$. If E stands for the expenses of the firm on all intermediate goods, the consumption of i is given by $q(i) = E/M_s\bar{p}$. Hence, we have

$$\begin{aligned} X_e &= M_s\left[\frac{E}{\alpha M_s\bar{p}}(1 + \log\beta) - \frac{E}{\alpha M_s\bar{p}}\log\frac{E}{\alpha M_s\bar{p}}\right] \\ &= \frac{E}{\alpha\bar{p}}(1 + \log\beta) + \frac{E}{\alpha\bar{p}}\log\frac{\alpha M_s\bar{p}}{E}, \end{aligned}$$

which is clearly increasing in M_s.

Each intermediate good is produced by one *specialized firm* under a constant marginal cost c (measured in terms of the numéraire) while using some fixed requirements of land S_s and of labor L_s. Hence, there is a continuum M_s of specialized firms.

The units of M_e and M_s are chosen for $S_e = S_s = 1$. Depending on the activities of the exporting firms, we may then have $M_e \geq M_s$. For example, if the exporting firms belong to the manufacturing sector, it is likely that $M_e > M_s$

because such firms use land extensively in comparison to service firms. By contrast, if the exporting firms are the headquarters of multinational firms, one expects $M_e < M_s$.

Following the same approach as the one taken in Section 7.2, we may write the profit function of an *exporting firm* located at x as follows:

$$\pi_e(x) = \int_X \{v[q(x, y)] - [p(y) + t|x - y|]q(x, y)\}m_s(y)dy \\ - R(x) - WL_e,$$

where $p(y)$ is the common price of the intermediates available at y, $q(x, y)$ the quantity of each intermediate good purchased by the firm at x from a specialized firm at y, t the common transportation rate of the intermediate goods, and $m_s(y)$ the number of specialized firms located at y. For the moment, the wage rate W is treated as a given constant within the urban area. Maximizing the term between curly brackets yields (7.5), and thus the profit function of an exporting firm at x becomes

$$\pi_e(x) = \int_X \beta m_s(y) \exp -\alpha[p(y) + t|x - y|]dy \\ - R(x) - WL_e. \tag{7.45}$$

Similarly, the profit function of a specialized firm at x is

$$\pi_s(x) = [p(x) - c] \int_X q(y, x)m_e(y)dy - R(x) - WL_s.$$

Using (7.5) again, we obtain

$$\pi_s(x) = [p(x) - c] \int_X \alpha\beta m_e(y) \exp - \alpha[p(y) + t|x - y|]dy \\ - R(x) - WL_s. \tag{7.46}$$

The common price of the intermediate goods is again

$$p_s^* \equiv p^*(x) = c + 1/\alpha,$$

which is to be interpreted along the lines of (7.6). Substituting this result in (7.45) and (7.46), we obtain

$$\pi_e(x) = \gamma \int_X m_s(y) \exp(-\alpha t|x - y|)dy - R(x) - WL_e \tag{7.47}$$

$$\pi_s(x) = \gamma \int_X m_e(y) \exp(-\alpha t|x - y|)dy - R(x) - WL_s, \tag{7.48}$$

where $\gamma \equiv \beta \exp -(\alpha c + 1)$.

Clearly, (7.47) and (7.48) are essentially identical to (7.7) and (7.8). Therefore, the analysis developed in Section 7.2 similarly applies so that

Proposition 7.1 holds under some obvious modifications ($N \to M_e$ and $M \to M_s$). In particular, when there are more specialized firms than exporting firms, the equilibrium involves the agglomeration of the exporting firms located with some specialized firms, whereas the remaining specialized firms sandwich the integrated district. This result sheds some light on the internal structure of CBDs: the headquarters of multinationals agglomerate with some service firms, forming the nucleus of the CBD, whereas other service firms locate in the outer ring of the CBD.

This model can be extended by adding a residential sector as well as a labor market. To this end, the equilibrium conditions of Section 6.3.3 must be generalized to the case of two types of firms (the exporting and specialized firms). It should then be clear to the reader that combining the results of Sections 6.3.3 and 7.2 allows one to show that the configurations presented in Figure 7.5 are possible equilibria in which H ≡ households, E ≡ exporting firms, and S ≡ specialized firms.

In Figure 7.5, diagram (a) describes a city in which the exporting firms form the nucleus of the CBD. Diagram (b) may depict an old industrial city in which the service firms agglomerate at the very center of the city, whereas factories are mostly in the outer ring of the business district. In diagram (c), in addition to the central complex formed by both specialized and exporting firms, the exporting

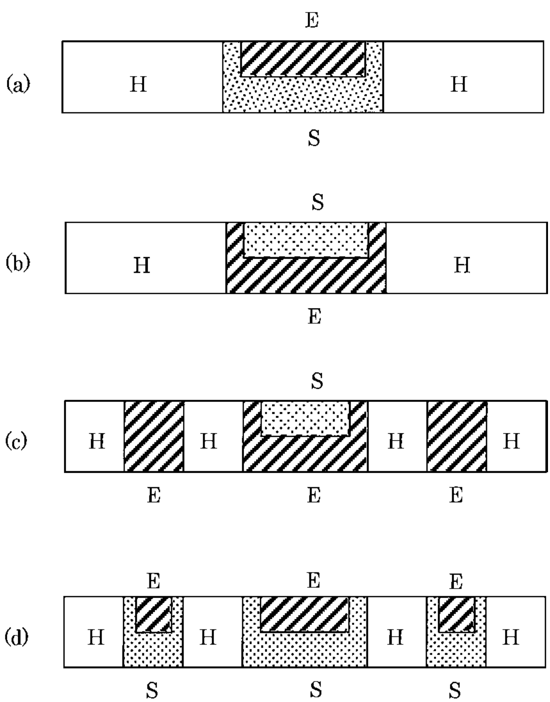


Figure 7.5: The equilibrium city configurations with exporting and specialized service firms.

firms also form two secondary employment centers because of their relatively high consumption of land; however, these firms still buy their intermediate goods from the CBD. In diagram (d), the entire metropolitan agglomeration involves three connected cities in which all the intermediates' services are supplied to the exporting firms in the three cities.

7.5.2 Formation of Secondary Employment Centers

In the previous section, firms were assumed to be "small" relative to the market size. This provides an incomplete picture of what we observe in the real world because cities are often molded (at least partially) by a few firms that are "large" relative to the urban labor market. Many examples may be found in the industrialized world in which medium-size cites welcome a large plant of a multinational firm (think of Toyota in Albany [Kentucky] or Renault in Flins [France], whereas NEC, one of the largest Japanese electronics firms, has decentralized its mass-production plant in medium-size cities throughout Japan). The entry of such firms dramatically affects the nature of competition on both the labor and land markets in the city. Furthermore, by the choice of their location, such firms may create secondary employment centers. Finally, their entry may also attract workers from other regions and cities, thus expanding the local population. These issues have not yet been considered so far.

To study the impact of large firms on the urban spatial structure, we consider a simple setting developed by Fujita, Thisse, and Zenou (1997). A large firm considers locating a new plant in a city where none of the existing businesses have a significant share of the urban labor market. The city itself is considered to be a small open economy subject to inflow migrations. Because the entrant is large relative to the city size, its location can therefore be viewed as a *secondary employment center*. By choosing a particular location, the firm affects the process of competition on the labor market in a rather complex way; it also affects the land rent, especially through the migration of rural workers who are hired by the firm. Because it is large relative to the urban market size, the entrant anticipates the impact of its location on the residential equilibrium owing to the migration of new workers.

Because the firm pays lower wages in the neighboring rural area where the land rent is lower, competition for workers generates a dispersion force that pushes the entrant away from the city center. However, as explained above, the existence of specialized firms located in the CBD, as well as the information flows generated by the CBD firms, acts as an agglomeration force that pulls the firm toward the city center. Consequently, the emergence of a secondary employment center appears as the outcome of the interplay of these two opposing forces.

To formalize this problem, we consider the now familiar model of the monocentric and linear city whose population size may increase with the migration of new workers. The existing firms are located in the CBD treated as a point and behave competitively on the labor market. In setting up its plant in the city, the new firm will compete with the incumbents in the labor market. If the entrant locates at the CBD, then it must compete with the existing firms to attract workers. This results in a higher wage that also pulls some workers from the neighboring rural areas who come in small numbers. If the firm locates at the outskirts of the city, its labor force is mainly constituted by new workers who choose to reside in the urban area, and, therefore, the firm brings new workers in the city in close proportion to its labor needs. In this case, the firm does not compete with the CBD firms and is able to offer a lower wage. This is possible because workers pay a lower rent on the land market and a lower commuting cost.

Consider a linear city in which all existing firms are small and located in the CBD, which is taken as the origin of the location space $X = (-\infty, \infty)$. In the spirit of the open-city model (see Chapter 3), the population N of workers is variable to allow for the migration of new workers. Workers are homogeneous, each consuming a fixed amount of land normalized to 1 and a variable amount of a composite good z (considered as the numéraire). Because the consumption of land is fixed, the workers' utility level can be expressed through the consumption z of the composite good. When they do not reside in the city, workers can guarantee themselves a reservation utility level equal to $\bar{z}$. Because the utility level decreases with the population size, workers will migrate to the city until they reach the utility level $\bar{z}$. Let 0 be the city center and y_e the location chosen by the entrant. Then, the budget constraint of a new or incumbent consumer at x working at y_i ($i = 0, e$) is as follows:

$$\bar{z} + t|x - y_i| + R(x) = w(y_i), \tag{7.49}$$

where t is the unit commuting cost, $R(x)$ the land rent at x, and $w(y_i)$ the wage rate at y_i. This implies that bid rent for the CBD workers is

$$\Psi(x; w) = w - \bar{z} - tx. \tag{7.50}$$

Before the entry of the new firm, the city is in the following state. If each worker in the city receives a wage w, the residential equilibrium condition at $x \in X$ is

$$\bar{z} + tx + R(x) = w.$$

When the opportunity cost of land is normalized to zero, this implies that the city expands on both sides up to the distance:

$$b = \frac{w - \bar{z}}{t}.$$

Consequently, the labor supply function is

$$N^{s}(w) = \frac{2(w - \bar{z})}{t}.$$

The optimizing behavior of the existing CBD firms is subsumed in the following linear labor demand function:

$$N^{d}(w) = \frac{A - w}{\theta},$$

where A and θ are positive constants. The market clearing wage is then

$$w^{*} = \frac{At + 2\bar{z}\theta}{t + 2\theta},$$

whereas the corresponding equilibrium employment is denoted by $N^{*} \equiv 2(A - \bar{z})/(t + 20)$. It is easy to check that the labor demand function can be rewritten as follows:

$$N^{d}(w) = N^{*} - \frac{w - w^{*}}{\theta}.$$

Given this monocentric city, we consider a new large firm, called e, entering the city. We assume this is a branch of a nationwide corporation that chooses the production target $\bar{Q}$ and product price $\bar{p}$ for its subsidiary as well as the region where it is to be located. The local manager selects a specific location within this region and the wage to be paid to the workers. The labor requirement of this firm is fixed and equal to $\bar{L}$, whereas its land consumption is assumed to be zero for simplicity. If firm e locates at y_{e} and pays a wage w_{e}, its profit function is

$$\Pi_{\mathrm{e}} = \bar{p}\bar{Q} - w_{\mathrm{e}}\bar{L} - (ay_{\mathrm{e}} + c)\bar{Q}, \tag{7.51}$$

where a stands for the accessibility cost to the CBD services per unit of output and distance and c is the marginal cost of all other inputs whose prices are assumed to be independent of the firm location. Everything else being equal, decreasing the distance to the CBD leads to a lower unit cost for the firm.

Because firm e is large relative to the city, it anticipates the impact of its location and wage choice on the corresponding markets. Three types of equilibrium configurations may arise, which are depicted in Figure 7.6.

Case 1. Firm e locates at the CBD (or near the CBD). Then, firm e competes with the CBD firms to attract its whole labor force. Specifically, the labor demand curve is shifted to the right by an amount equal to $\bar{L}$, whereas the supply of labor is given by the existing population N^{*} augmented by some immigrants. The land rent is uniformly shifted upward (see Fig. 7.6 (a)). The new equilibrium wage is obtained by solving the labor market clearing condition:

$$N^{d}(w) + \bar{L} = N^{s}(w),$$

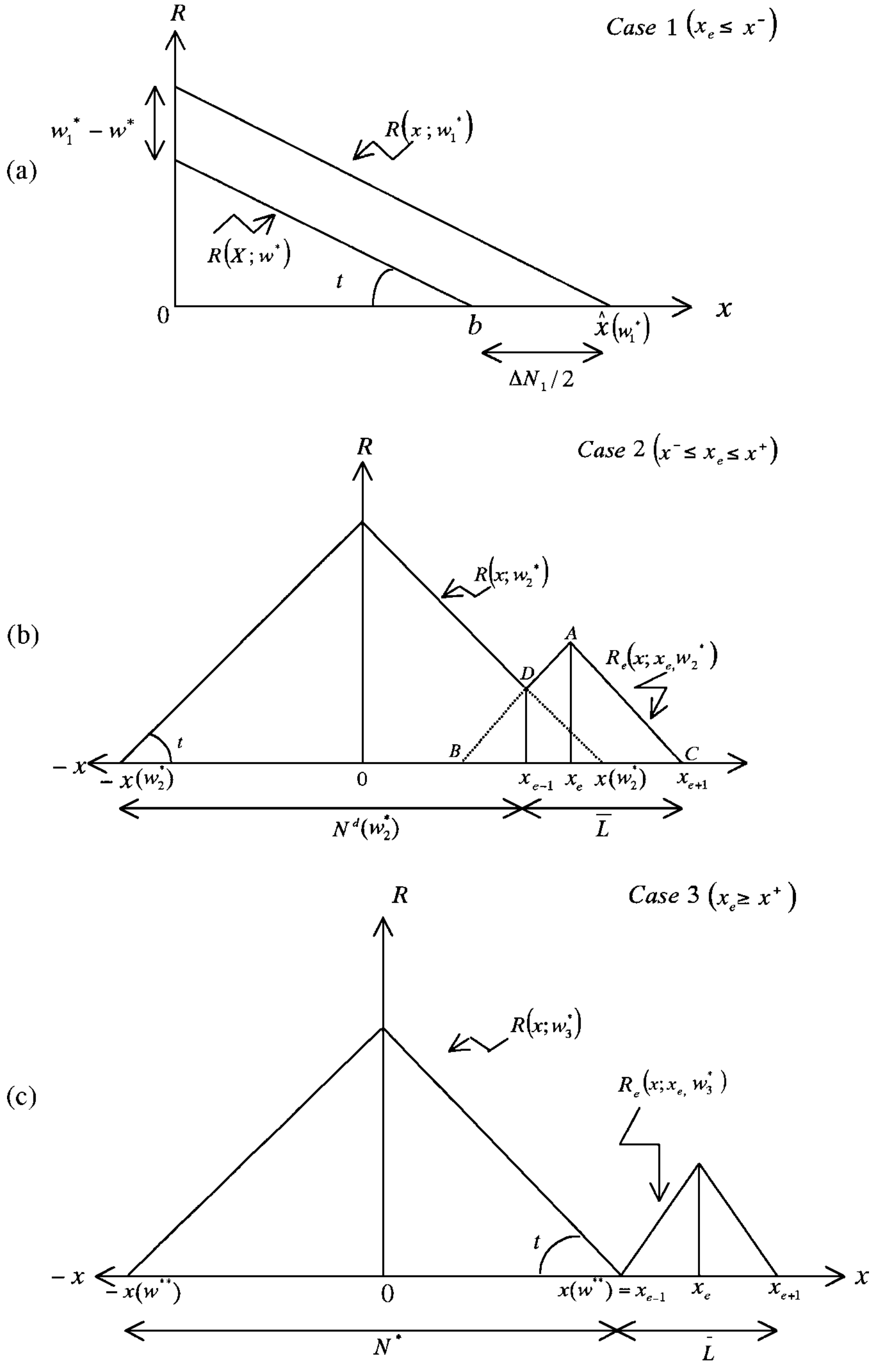


Figure 7.6: The formation of secondary employment centers.

that is,

$$w_1^* = w^* + \frac{\bar{L}}{1/\theta + 2/t}. \tag{7.52}$$

In this case, the labor market is fully integrated because all people work in (or near) the CBD; the city employment increases by

$$\Delta N_1 = \frac{\bar{L}}{1 + t/2\theta},$$

which is always smaller than $\bar{L}$.

The same results hold true if firm *e* locates at y^- smaller than

$$x^- \equiv \frac{N^*}{2} + \frac{\bar{L}}{t/\theta + 2} - \bar{L}$$

because the rise in population keeps the urban pattern symmetric.

Case 2. Firm *e* locates between x^- and x^+ where

$$x^+ = (N^* + \bar{N})/2.$$

The firm still competes with the CBD firms by attracting some workers who used to work for them (because $y_e < x^+$). However, its labor force is also formed by immigrants (because $y_e > x^-$) so that it has some monopsony power on its labor market. This implies that the urban labor market is segmented between the central market and the market formed around firm *e*, which is reflected in the two peaks of the land rent curve (see Fig. 7.6 (b)). The employees of firm *e* reside near the city boundary, thus allowing this firm to pay a lower wage owing to the lower commuting costs. The bid rent for the consumers working in the new firm is

$$\Psi(x; y_e, w(y_e)) = w(y_e) - \bar{z} - t|x - y_e|, \tag{7.53}$$

where $w(y_e)$ is the wage paid by the entrant.

Because the bid rents (7.50) and (7.53) must be the same at the labor-market boundary, the equilibrium wage at the CBD is

$$w_2^*(0) = w_1^* - A(\theta)(y_e - x^-),$$

which obviously decreases when the entrant locates farther away from the CBD, whereas the wage set by the entrant is

$$w_2^*(y_e) = \frac{w^* + 2\bar{z} + 2t\bar{L}}{3} + \frac{A(\theta)t(N^* + \bar{L})}{6} - \frac{[1 + A(\theta)]t}{3} y_e, \tag{7.54}$$

where $A(\theta) \equiv 1/(2 + 3t/2\theta)$. The urban employment rises by an amount $\Delta N_2(y_e)$, which is always larger than $\Delta N_1(y_e)$ but smaller than $\bar{L}$. The creation of a secondary employment center breaks the symmetry of the urban pattern.

Case 3. The new firm establishes itself outside the city at a distance $\bar{L}/2$ from the initial city fringe b. It hires immigrants only, whereas all former residents keep working for the CBD firms (see Fig 7.6 (c)). There is no competition between the two local labor markets, and thus both the preentry land rent and wage still prevail after entry in the initial urban area. Firm e is a monopsonist offering a wage given by

$$w_3^* = \bar{z} + \frac{t\bar{L}}{2}. \tag{7.55}$$

Urban employment is augmented by exactly $\bar{L}$, and a new center is created that interacts with the preexisting city only through firms located at the CBD.

To sum up: using (7.52)–(7.55), we see the wage rate paid by the entrant decreases as its location moves away from the CBD. This is exactly the dispersion force described in the foregoing.

Given the profit function (7.51), the entrant must balance the advantages provided by the greater proximity to the center against the intensity of competition on the labor market as reflected by a downward-sloping wage gradient.

Which pattern will arise depends on the relative intensity of these two opposing forces. Because the profit function is piecewise linear in y_e, the equilibrium location is always at one extremity of the feasible interval associated with each case. It is then easy to show the following:

Proposition 7.5 *When the profit function is given by (7.51), the equilibrium location of the entrant is such that*

1. *if* $a\bar{Q} > t\bar{L}$, *then* $y_e^* = 0$
2. *if* $t\bar{L} \geq a\bar{Q} > t\bar{L}[1 + A(\theta)]/3$, *then* $y_e^* = x^-$
3. *if* $t\bar{L}[1 + A(\theta)]/3 \geq a\bar{Q}$, *then* $y_e^* = x^+$.

When the intensity of communication with the service firms at the CBD is strong enough (Case 1), the new firm will locate together with its competitors on the labor market. At the other extreme, when the intensity of communication is low, the firm will choose to create an edge city at the fringe of the existing city by appealing to immigrants only (Case 3). In between, there is interdependence between the two local labor markets (Case 2). When the distance to the CBD rises, the interdependence between these two markets weakens smoothly, as shown by the fact that the firm offers a continuously decreasing wage. It is in this case that the entry of a new firm gives rise to the creation of a secondary employment center within the city. The location of this center with respect to the CBD eventually depends on the basic parameters of the urban economy. In particular, when θ increases, labor demand by CBD firms becomes less elastic, and thus firm e faces a fiercer competition on the city labor market. As a result, it tends to locate toward the peripheral location x^{+}.[28]

7.6 CONCLUDING REMARKS

One major conclusion of this chapter is that the clustering of firms and the emergence of local labor markets within a city may correspond to equilibrium outcomes once it is admitted that markets are imperfectly competitive. Although externalities are likely to be crucial in the formation of such agglomerations in the real world, we believe that the results presented here support our claim that the presence of imperfectly competitive markets is another major reason for the existence of these agglomerations.

Even though we have used fairly specific models, some general principles seem to emerge from our analysis. In particular, a commercial (employment) center is likely to form when products (intermediate goods) are differentiated enough, transport costs are sufficiently low, or both. This is in accord with several of the results derived in Chapter 6 for which we assumed externalities to be at work in an otherwise competitive environment. Interestingly, in Part III we will encounter again the same kind of logic in the formation of a core–periphery structure. In fact, the spatial monocentric structure obtained in Propositions 6.2, 7.2, and 7.3 can be interpreted as the urban counterpart of the core–periphery structure. We also notice that principles governing the working of spatial product markets are similar, mutatis mutandis, to those to be applied to spatial intermediate goods markets.

More surprising, perhaps, is that, when agglomeration occurs, it is often socially desirable. This is because consumers like to try each variety and, therefore, benefit from a spatial concentration of sellers at the center of the urban area. The need for agglomeration is even stronger when consumers have incomplete information about varieties and prices, for the grouping of firms allows a substantial reduction in consumers' search costs. Thus, besides the positive effects that agglomeration may have for firms (as discussed in Chapter 6), we have stressed here some of the positive effects of agglomeration for consumers/workers. However, we would certainly be the last ones to claim that agglomeration is always optimal. Quite the reverse is true. We have identified some major dispersion forces that may pull typical urban economic activities away from the city center. In addition, even when agglomerating firms is socially desirable, we have seen that the place selected by firms may not be socially optimal. We will return to this problem later.

APPENDIX

A. Consider first any residential section $[a_1, a_2]$ in the urban area $[-l, l]$. From (7.9), we have

$$\frac{d^2\Psi(x)}{dx^2} = \gamma\alpha^2 t^2 \left[\int_{-l}^{a_1} m(y)\exp{-\alpha t(x-y)}dy + \int_{a_2}^{l} m(y)\exp{-\alpha t(y-x)}dy\right],$$

for $m(y) = 0$ over $[a_1, a_2]$. Because this expression is strictly positive, $\Psi(x)$ is strictly convex over $[a_1, a_2]$.

Assume now that $[a_1, a_2]$ is a business section. Taking again the second derivative of (7.9) yields

$$\begin{aligned}\frac{d^2\Psi(x)}{dx^2} &= -\gamma\alpha t[\exp{-\alpha t(x-a_1)} + \exp{-\alpha t(a_2-x)}]\\ &\quad + \gamma\alpha^2 t^2\left[\int_{-l}^{a_1} m(y)\exp{-\alpha t(x-y)}dy + \int_{a_2}^{l} m(y)\exp{-\alpha t(y-x)}dy\right]\\ &\leq -\gamma\alpha t[\exp{-\alpha t(x-a_1)} + \exp{-\alpha t(a_2-x)}]\\ &\quad + \gamma\alpha^2 t^2\left[\int_{-l}^{a_1} \exp{-\alpha t(x-y)}dy + \int_{a_2}^{l} \exp{-\alpha t(y-x)}dy\right]\\ &= -\gamma\alpha t[\exp{-\alpha t(x+l)} + \exp{-\alpha t(l-x)}],\end{aligned}$$

which is always negative, and thus $\Psi(x)$ is strictly concave over $[a_1, a_2]$.

A similar argument applied to (7.10) shows that $\Phi(x)$ is strictly convex (concave) over any business (residential) section.

B. Assume that firm 1 is located at $y_1 < l/2$ and quotes a price p_1, whereas the others are at $l/2$, and set $p^* = c + M/\alpha(M-1)$. Firm 1's profit is given by

$$\begin{aligned}\pi_1(p_1, y_1) = (p_1 - c)\Bigg[&\frac{y_1}{1+(M-1)\exp(\Lambda-\Theta)} + \frac{l}{2} - y_1\\ &- \frac{1}{\alpha\tau}\ln\frac{1+(M-1)\exp(\Lambda+\Theta)}{1+(M-1)\exp(\Lambda-\Theta)}\\ &+ \frac{l/2}{1+(M-1)\exp(\Lambda-\Theta)}\Bigg],\end{aligned}$$

where $\Theta \equiv \alpha t/(l/2 - y_1)$ and $\Lambda \equiv \alpha(p_1 - p^*)$. Therefore,

$$\begin{aligned}\operatorname{sign}\frac{\partial \pi_1}{\partial x_1} = \operatorname{sign}\Bigg\{&tle^{\Theta}[e^{\Theta} + (n-1)e^{\Lambda}]^2 + \frac{1}{\alpha}(e^{2\Theta}-1)[e^{\Theta} + (n-1)e^{\Lambda}]\\ &\times[1+(n-1)e^{\Theta}e^{\Lambda}] - 2ty_1e^{\Theta}[1+(n-1)e^{\Theta}e^{\Lambda}]\Bigg\}.\end{aligned}$$

If $\alpha tl \leq 2$, a lower bound for the right-hand side of this expression can be obtained by replacing $1/\alpha$ by $tl(1-2/n)/2$ and is given by

$$\frac{tl}{2}\left\{e^{\Theta}\left[1+e^{2\Theta}\left(1-\frac{4y_1}{l}\right)\right](n-1)^2 e^{2\Lambda}\right.$$

$$\left[(e^{4\Theta}-1)+4e^{2\Theta}\left(1-\frac{2y_1}{l}\right)\right](n-1)e^{\Lambda}$$

$$\left.+e^{\Theta}\left[3(e^{2\Theta}-1)+2\left(1-\frac{2y_1}{l}\right)\right]\right\}.$$

This expression is always positive for $y_1 < l/2$ regardless of Λ when $\alpha tl \leq 2$, and thus firm 1 wants to join the others at $l/2$.

It remains to show that it is more profitable for firm 1 located at $l/2$ to charge p^*. Its profit is $\pi_1 = (p_1 - c)l\boldsymbol{P}_1$, where

$$\boldsymbol{P}_1 = \frac{\exp(-\alpha p_1)}{\exp(-\alpha p_1) + (M-1)\exp(-\alpha p^*)}.$$

It can then readily be verified that

$$\frac{\partial \pi_1}{\partial p_1} = (p_1 - c)\alpha l\boldsymbol{P}_1(\boldsymbol{P}_1 - 1) + l\boldsymbol{P}_1$$

and

$$\frac{\partial^2 \pi_1}{\partial p_1^2} = (p_1 - c)\alpha^2 l\boldsymbol{P}_1(\boldsymbol{P}_1 - 1)(2\boldsymbol{P}_1 - 1) + 2\alpha l\boldsymbol{P}_1(\boldsymbol{P}_1 - 1).$$

Evaluating $\partial^2\pi_1/\partial p_1^2$ at any price for which $\partial\pi_1/\partial p_1 = 0$ gives $-\alpha l\boldsymbol{P}_1 < 0$. The profit function π_1 is therefore strictly quasi-concave, and thus the solution $p_1 = p^*$ of $\partial\pi_1/\partial p_1 = 0$ is a profit-maximizing price when firm 1 is with the others.

C. Let T_C be the expected transport costs when the consumer starts his or her search in the cluster and T_I the same cost when the search starts at the isolated firm. First, we have the following inequality:

$$\begin{aligned} T_C \leq{} & [t_0 + t|x - y_C|] + [(t_0 + t|x - y_C|)[1 - (1 - D_I)(1 - D_C)^{M-2}]] \\ & + [(t_0 + t\Delta)(1 - D_I)(1 - D_C)^{M-2}] \\ & + [(t_0 + t|x - y_I|)[(1 - D_I)(1 - D_C)^{M-2} \\ & - (1 - D_I)(1 - D_C)^{M-1}]] \\ & + [(t_0 + t\Delta + t_0 + t|x - y_C|)(1 - D_I)(1 - D_C)^{M-1}], \end{aligned} \tag{A.1}$$

where

1. the first (bracketed) term, $t_0 + t|x - y_C|$, is the cost of going to the cluster;

2. the second term corresponds to the cost of returning home, $t_0 + t|x - y_{\mathrm{C}}|$, times the probability $1 - (1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-2}$ that the consumer assigns to buying in the cluster;
3. the third term stands for the cost of traveling from the cluster to the isolated firm $t_0 + t\Delta$ times the probability $(1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-2}$ the consumer wants to go there from the cluster;
4. the fourth term represents the cost to go home from the isolated firm, $(t_0 + t|x - y_{\mathrm{I}}|)$, weighted by the probability $(1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-2} - (1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-1}$ that the consumer does not want to return to the cluster;
5. the last term is the cost of going back to the cluster before returning home, $t_0 + t\Delta + t_0 + t|x - y_{\mathrm{C}}|$, weighted by $(1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-1}$.

The inequality holds in (A.1) because the last weight, $(1 - D_{\mathrm{I}})(1 - D_{\mathrm{C}})^{M-1}$, overestimates the probability that the consumer wants to return to the cluster, thus making the last term larger than it should be.

Second, we also have

$$\begin{aligned} T_{\mathrm{I}} \geq {} & [t_0 + t|x - y_{\mathrm{I}}|] + [(t_0 + t|x - y_{\mathrm{I}}|)D_{\mathrm{I}}] \\ & + [(t_0 + t\Delta)(1 - D_{\mathrm{I}})] + [(t_0 + t|x - y_{\mathrm{C}}|)(1 - D_{\mathrm{I}})], \end{aligned} \tag{A.2}$$

where

1. the first term, $t_0 + t|x - y_{\mathrm{I}}|$, is the cost of traveling to the isolated firm;
2. the second term is the cost of returning home, $t_0 + t|x - y_{\mathrm{I}}|$, times the probability D_{I} of buying from this firm;
3. the third term stands for the cost of going to the cluster from the isolated firm, $t_0 + t\Delta$, weighted by the probability, $1 - D_{\mathrm{I}}$, the consumer assigns to this event; and
4. the last term represents the cost of going home from the cluster, $t_0 + t|x - y_{\mathrm{C}}|$, times the probability of being in the cluster, $1 - D_{\mathrm{I}}$.

The inequality holds in (A.2) because we do not account for the possibility that the consumer might want to return to the isolated firm after having visited the cluster.

Subtracting (A.1) from (A.2) and using the triangle inequality, we obtain after some manipulations

$$\begin{aligned} T_{\mathrm{I}} - T_{\mathrm{C}} & \geq -2t\Delta + (1 - D_{\mathrm{I}})[t_0 + 2t\Delta - (1 - D_{\mathrm{C}})^{M-2}t_0 \\ & \qquad - (t_0 + 2t\Delta)(1 - D_{\mathrm{C}})^{M-1}] \\ & \equiv f(\Delta). \end{aligned}$$

Let $\Delta = 0$ and choose $\hat{M}$ large enough for $(1 - D_{\mathrm{C}})^{M-1}$ to be smaller than $1/2$. Clearly, $f(0) > 0$ because $t_0 > 0$ and $D_{\mathrm{I}} < 1$, for otherwise a consumer would visit a single place, that is, the cluster, and the problem would become trivial. On the other hand, $f(\Delta) < 0$ when Δ is large enough. Because $f(\Delta)$ is continuous, the intermediate value theorem implies a value $\Delta(M)$ exists such that $T_I - T_C > 0$ for $0 \le \Delta \le \Delta(M)$.

Because $\Delta(M)$ is the smallest solution of the equation

$$-2t\Delta + (1 - D_I)[t_0 + 2t\Delta - (1 - D_C)^{M-2}t_0 - (t_0 + 2t\Delta)(1 - D_C)^{M-1}] = 0,$$

it can readily be verified that $\Delta(M)$ increases with M.

NOTES

1. As discussed by Martinez-Giralt and Neven (1988), price competition may be so aggressive that each firm chooses to concentrate its own production into a single plant located away from its rivals.
2. Observe that adding a free-entry condition would allow us to determine the number M of firms by the zero profit condition. This will be done later on.
3. Note that the entropy-type utility function is a close relative of the CES (see Anderson et al. 1992, chaps. 3 and 4, for detailed analyses).
4. Observe that any equilibrium pattern can be shown to be symmetric (Fujita 1986).
5. Observe that Papageorgiou and Thisse (1985) have also shown the emergnce of the agglomeration of firms and households in a related model but with a finite number of locations.
6. This is not be necessarily true with a concave demand field in which consumers do not care about small increases in distance in the vicinity of $x = 0$; see Fujita (1986) for a counterexample.
7. The latter result confirms what we have seen in the previous chapter.
8. More details can be found in Fujita (1986; 1988).
9. Unless every transaction must occur at given marketplaces, as in the Thünian model (see Chapter 3).
10. This is done in Fujita (1988) for linear demand functions. However, his analysis carries over to the case of exponential demand functions considered here.
11. Contrary to a widespread opinion, this result is not driven by the existence of boundaries. To see this, consider a continuous distribution over the real line. Once the two firms locate back to back at the median of the distribution. We believe that several of the results presented here could be extended to this framework.
12. See Netz and Taylor (2001) for some empirical evidence.
13. This assumption has been relaxed by De Fraja and Norman (1993), who use linear demands derived from a quadratic utility function.
14. Because the total consumption of the differentiated product is constant and the same across consumers, we may set $\beta = 1$ in (7.4). Without loss of generality, the utility U is multiplied by α. It is then easy to see that (7.4) and (7.38) are equivalent. See Anderson, de Palma, and Thisse (1992, Proposition 3.7) for more details.
15. See, for example, Anderson et al. (1992, chap. 3).
16. This model is often given a probabilistic interpretation, which can be found in Anderson et al. (1992, chap. 9). In this case, it is assumed that consumers are

influenced by various tangible as well as intangible factors at the moment of their choice and that the relative importance of these factors may change owing to external factors. This implies that consumers' purchase decisions are not based solely on the full prices but also on firm-specific factors that are typically perceived differently by different consumers. Such a behavior means that consumers at the same location do not react in the same way to a firm's unilateral change in strategy. The presumably wide array of factors influencing consumers' shopping behavior makes it problematic for a firm to predict exactly a consumer's reactions to a reduction in price. In other words, the firm assigns a probability between 0 and 1 to whether a particular consumer on a particular date will respond to a price difference by switching firms. This is modeled by assuming that consumers maximize a *random utility*. Firms thus implicitly sell heterogeneous products, and the random term in the consumer's utility expresses his matching with firms at the time of purchase. Even if prices do not vary, consumers do not always purchase from the same firm over time. The values of the choice probabilities $P_i(x)$ depend on those of the full prices: the higher the latter, the lower the former. Consequently, consumers' behavior encapsulate a tendency to buy from the cheapest shops. The expected demand to firm i is equal to the integral of the choice probabilities over the market space; it is continuous in prices and locations when μ is strictly positive. However, the continuity of profits does not suffice to restore the existence of an equilibrium. Additional restrictions on the parameters are necessary: the relative importance of the transport costs must be small compared with that of the idiosyncratic components of the individual preferences.

17. Broadly speaking, this branch of economic geography aims at analyzing and forecasting any movement over space that results from human action (see, e.g., Anas 1987; Sen and Smith 1995; Webber 1979). Regarding gravity models, two additional remarks are worth making. First, the empirical relevance of gravity models was known in the economics profession for a long time; see, for example, Isard (1956) and Tinbergen (1962). Second, Anas (1983) has shown that gravity- and logit-type models can be derived from the maximization of a random utility, thus casting them into the realm of economic theory.
18. For a proof, see Anderson et al. (1992, chap. 9).
19. Here, consumers' locations are fixed; hence, there is no mutual attraction through the interplay of consumers' and firms' locations. Introducing a land market with elastic demand for land would strengthen the tendency toward agglomeration (see, e.g., Fujita and Thisse 1991).
20. See also Ben-Akiva, de Palma, and Thisse (1989) and De Fraja and Norman (1993) for similar results when consumers have alternative standard preferences for differentiated products. We may then safely conclude that the existence of an agglomerated equilibrium under sufficient product differentiation is robust against alternative specifications of demand.
21. The reader must remember that Hotelling considered a two-stage game in which firms choose first locations and then prices. What we have just seen remains basically the same for such a game (Anderson et al. 1992, chap. 9).
22. This result had already been anticipated by Launhardt ([1885] 1993, 150, of the English translation), for whom "the most effective of all protective tariffs [is] the protection through poor roads."

23. This is confirmed by various papers in industrial organization studying product selection in multicharacteristic spaces. In this respect, the main result is from Irmen and Thisse (1998), who have shown that, in equilibrium, firms maximize differentiation in the "dominant" characteristic but minimize differentiation in all others. Consequently, when transport costs are no longer the dominant characteristic, firms seek their profit margins in differentiated products instead of geographical separation. Indeed, once products are differentiated and transport costs low, firms are able to benefit from a better accessibility to the whole market by locating at the market center without being harmed by tough price competition.
24. Greenhut (1981) has presented the results of a survey of 241 firms in Germany, Japan, and the United States. Mill pricing is used by 29% of the firms sampled, whereas the remaining firms follow various types of delivered pricing. See also Phlips (1983, chap. 1). Shopping and shipping models in spatial competition correspond to the standard distinction between integrated and segmented markets in international trade.
25. The model can be extended to the case in which consumers are also unsure about firms' prices when these are endogenous.
26. One point needs clarification. We consider a static model in which consumers shop only once. This is of course a caricature of reality. Instead, one should think of an environment in which new consumers arise over time in different locations. Another interpretation is that shops change their goods from time to time, as in the case of fashion stores.
27. This is apparent in Stahl (1982), who has shown that the noncooperative market outcome does not belong to the core of the cooperative game, the players of which are the firms.
28. The model can be extended to the case in which the entrant is able to capitalize the land rent it creates. The results are qualitatively the same, though there is a stronger tendency for the entrant to locate in the periphery (Fujita et al. 1997).

PART III

FACTOR MOBILITY AND INDUSTRIAL LOCATION

8

Industrial Agglomeration under Marshallian Externalities

8.1 INTRODUCTION

The work of Marshall has been very influential in regional and urban studies and has led to the concept of *Marshallian externalities*, which aim at accounting for the benefits associated with cluster formation. As seen in Chapter 1, these benefits arise because of

1. the formation of a highly specialized labor force and the development of new ideas, both based on the accumulation of human capital and face-to-face communications (the latter has been analyzed in Chapter 6);
2. the availability of specialized input services (see Chapter 4); and
3. the existence of modern infrastructure (see Chapter 5).[1]

From the point of view of this chapter, the main distinctive feature of these externalities is that they affect only the agents belonging to the same geographical area. They do not spread over other regions or, more precisely, their impact on distant regions may be considered negligible.[2]

The now standard classification of Marshallian externalities is attributed to Hoover (1936, chap. 6): (1) the *localization economies*, which are defined as the benefits generated by the proximity of firms producing similar goods; and (2) the *urbanization economies*, which are defined by all the advantages associated with the overall level of activity prevailing in a particular area. These external effects have been studied extensively from the empirical standpoint (see, e.g., Henderson 1988, chap. 5, and the references therein).[3] For example, Hanson (1996, 1266) accurately points out that "the fact that New York City remains a major apparel producer is perhaps the most persuasive evidence one can find of localization economies."

The evidence shows that both types of external effects are at the origin of several specialized and prosperous areas. According to Jacobs (1969), urbanization economies are predominant, whereas Porter (1998, chap. 7) has argued that the main reason for the success of industrial clusters in the global economy is the presence of strong localization economies. The conclusion to the debate over localization versus urbanization economies is an empirical one and, for the time being, it seems hard to foresee what the outcome will be. However, one may conjecture that the results are likely to depend on the nature of the industry as well as on the size of the area in question.

To a large extent, the idea of localization economies also explains the growth and success of *industrial districts*, that is, regions that accommodate many small firms producing similar goods and that benefit from the localized accumulation of skills associated with workers residing in these places (Becattini 1990). Some industrial districts are engaged in high-tech activities (Saxenian 1994), but others are involved in more traditional, labor-intensive activities, many of which can be found in the "Third Italy" (Pyke et al. 1990, chaps. 4 and 5): Sassuolo specializes in ceramic tiles, Prato is known for textiles, shoes are made in Montegranaro, and wooden furniture in Nogara.

In any case, the combination of various factors turns out to be essential for the localized accumulation of various types of knowledge within a region.[4] For example, Feldman and Florida (1994, 226) concluded a detailed study of the geography of innovation in the United States by noticing that

> innovation is no longer the province of the inventor, the risk-taking entrepreneur, the insightful venture capitalist, or the large resource-rich corporation. Innovation instead has its sources in a broader social and spatial structure – a landscape of agglomerated and synergistic social and economic institutions welded into a technological infrastructure for innovation.

The foregoing discussion as well as early work in urban economics (Henderson 1988, chap. 5) leads us to consider Marshallian externalities as factors positively affecting the local productivity through the accumulation of some particular inputs available in the same area. Ever since the pioneering works of Thünen ([1826], 1966), Marshall ([1890], 1920), and Weber ([1909], 1929), such external effects have been at the heart of most explanations of economic agglomerations, but they tend to be invoked without much being said about their underpinnings. The analyses provided in previous chapters should shed light on what lies at the origin of such external economies. In this chapter, our research strategy is to view Marshallian externalities as technological externalities without accounting for the details of their specificity. This shortcut allows us to explore the common implications of Marshallian externalities for the spatial distribution of production activities.

The purpose of this chapter is twofold. First, we want to show how the introduction of externalities into the neoclassical theory of regional economics has affected the convergence property. We will see that even a small departure from the standard framework is sufficient for the most typical market outcome to involve a core–periphery structure in which one region is more prosperous than the other. Stated differently, there is no longer convergence between nations or regions. Second, we wish to discuss the optimal strategies of firms facing the possibility of enjoying lower costs inside a cluster or of establishing their plants in spatially separated markets in order to enjoy more market power. The main result here is the emergence of an asymmetric distribution of firms in an otherwise symmetric environment.

In Section 8.2, we consider a simple setting with one mobile factor and one immobile factor developed by Michel, Perrot, and Thisse (1996). As we know, the neoclassical model predicts that the mobility of one production factor suffices to yield regional convergence. The situation may change dramatically if there is a production externality whose local intensity depends on the quantity of the mobile factor gathered into one region. We will see that such an externality may lead to *the expansion in wealth of the region* with the initial advantage at the expense of the other, thus providing an example of the circular causation process proposed by Myrdal (1957, chap. 3) to explain economic development. Similar results will be obtained in more specific contexts in Chapter 9.[5]

In the next two sections, we take a similar line of reasoning but focus on a microeconomic setting in which firms compete in price while being able to benefit from Marshallian externalities in their location choices (Belleflamme, Picard, and Thisse 2000). Specifically, we follow Chipman (1970) by assuming that firms belonging to the same sector benefit from a higher productivity when they locate together. We want to highlight how these external effects interact together with the dispersion forces generated by market competition within the global economy to lead to cluster formations of different sizes. Indeed, if localization economies are obviously an agglomeration force, it is also well known that geographical proximity renders price competition on the product market fiercer (Section 7.3). In other words, firms have an incentive to separate from one another to enjoy local market power. Consequently, whereas firms enjoy low costs when they concur in their locational choices, they can sell their products at higher prices when they are dispersed.

This is not the end of the story, however. Even if price competition is relaxed through product differentiation, it is still true that firms want to be far apart when transport costs are high. Because the spatial distribution of demand is supposed to be unaffected by the locations and sizes of clusters, the cost reduction associated with the agglomeration may be more than offset by the fall in exports. By contrast, firms could enjoy higher profits by being local monopolists. Consequently, transport costs have to be low for firms to congregate. In other words, firms must be able to serve almost all markets equally (globalization) to enjoy the local advantages associated with the formation of a cluster (localization). Consequently, the formation of an industrial cluster appears to depend on the relative strength of three distinct forces: the magnitude of localization economies, the intensity of price competition, and the level of transport costs. In such a context, the number of firms is likely to matter, and we therefore consider two very different market structures, that is, competition within a *small group* (Section 8.3) and within a *large group* of firms (Section 8.4).

It is reasonable to believe that the intensity of localization economies varies across regions. This is why we consider in Section 8.3 an oligopoly in which the production cost reduction firms may enjoy by being together changes with the region in which they set up. Such a setting provides a benchmark to study the decisions of a small number of large firms behaving strategically because each firm is aware that its locational choice affects not only its production cost but also its rivals'. This setting also captures some critical elements of the trade-off faced by large firms in designing their competitive strategy such as those belonging to the German chemical clusters.

In Section 8.4, we move to the case of a large group of firms in which there is no explicit strategic interaction: each firm has a negligible impact on the others but is aware that its locational choice affects its production cost because this factor depends on where its competitors are located. This market structure will allow us to study the emergence of industrial clusters involving a large number of small firms, such as those existing in Italy.

8.2 FACTOR MOBILITY AND AGGLOMERATION ECONOMIES

Consider an economy formed by two regions $r = A, B$. There are one product and two production factors. The product is costless to transport. One factor is immobile, whereas the other responds to market disequilibrium by moving into the more attractive region. Although other interpretations are possible, it is convenient to think of the mobile factor as the *skilled* workers and the immobile one as the *unskilled* workers. Indeed, empirical studies suggest that skilled workers are more mobile than the unskilled between distant locations (e.g., Shields and Shields 1989; SOPEMI 1998) perhaps because education generates human capital that is easily transferable to other regions.[6]

In region r, the output Y_r is produced according to the production function

$$Y_r = E(H_r)F(H_r, L_r),$$

where H_r stands for the number of skilled in region r, with $H = H_A + H_B$, whereas L_r represents the number of unskilled with $L = L_A + L_B$. Both H and L are fixed. The total number of workers in region r is denoted by $P_r = H_r + L_r$. Because we wish to focus on the pure effects of the forces at play, we consider the case of symmetric regions, and thus $L_A = L_B = 1$.

The production function, common to the two regions, consists of two parts: (1) a neoclassical production function $F(H_r, L_r)$ with constant returns and diminishing marginal product and (2) an externality function $E(H_r)$ that has the nature of a Hicksian shift factor (as in Section 4.4). For notational simplicity, we set

$$F(H_r, 1) \equiv f(H_r),$$

where $f'(H_r) > 0$, $f''(H_r) < 0$, whereas the Inada condition $f'(0) = \infty$ is supposed to hold. In the present context, it seems natural to assume that the intensity of the externality increases with the number (or mass) of skilled residing in the same region, and thus E is continuously increasing on $[0, H]$. Such a production externality is a force fostering the agglomeration of the skilled into a single region. However, increasing the number of skilled in a given region decreases their marginal productivity gross of the externality effect. As a result, the wage the skilled workers earn in the large region may eventually decline, thus leading to a dispersed pattern of production.

The regional output markets are perfectly competitive. There are no transportation costs, therefore, the price of the output is the same in the two regions and is normalized to 1. Similarly, regional factor markets are perfectly competitive, and thus the regional wage of each factor is given by its regional marginal productivity. Because each firm in region r is negligible to the market, it considers the externality $E(H_r)$ as given. Therefore, since both regional markets for the skilled are perfectly competitive, we have

$$w_r^H = E(H_r)f'(H_r) \quad r = A, B. \tag{8.1}$$

Workers have identical preferences. The utility of a j-worker ($j = H, L$) living in region r ($r = A, B$) is given by

$$U_r^j = u\left(w_r^j\right) + e_r(P_r),$$

where $u(w_r^j)$ is the indirect utility derived from the wage w_r^j, and $e_r(P_r)$ is a consumption externality depending on the total population residing in region r.[7] We assume that the consumption externality has the same functional form across the two groups of workers and is given by

$$e_r(P_r) = v\left(P_r, \frac{P_r}{S_r}\right),$$

where P_r/S_r is the density of population in region r, and S_r the area of this region. Note that S_r need not be the physical size of the region; for example, S_r could be the given amount of any public infrastructure or natural amenity consumed by workers. As for the unskilled, we assume that the two regions have identical endowments: $S_A = S_B = S$.

The function v is the same across regions so that none of them has a structural advantage. We also assume that $v_1' > 0$ and $v_2' < 0$. The first inequality ($v_1' > 0$) expresses a *conviviality effect* of the type analyzed in Section 6.2: the larger the population in region r, the higher the potential for social interaction. However, as discussed in Chapters 3 and 5, the positive effect of a larger population in region r is counterbalanced by *crowding effects* expressed by the second inequality ($v_2' < 0$). For example, a larger population living in a given area is likely to face a higher pollution and crime rate that reduces the workers' welfare.

Hence, there is a trade-off because the first effect encourages the agglomeration of skilled workers, whereas the second fosters their dispersion. When additional insights are required, we use the following specification for the consumption externality:

$$v(P_r) = aP_r - b(P_r/S)^2, \tag{8.2}$$

where a and b are two positive constants expressing the relative importance of the amenity and crowding effects in workers' well-being.

Let us now describe the migration behavior of the skilled. Migration is driven by the *utility differential*:

$$\dot{H}_A = \left[u\left(w_A^H\right) + e(P_A)\right] - \left[u\left(w_B^H\right) + e(P_B)\right], \tag{8.3}$$

for, as will be seen, H_A and H_B are always positive. In this case, the utility a skilled worker obtains in a given region depends on the regional wage, which varies with the number of skilled workers but also with the various consumption externalities generated inside this region by the local population. Various types of externalities are thus operative.

When regional amenities are weakly affected by the migration of the skilled, it is convenient to assume that they are the same in both regions, in which case migration is governed by the *wage differential*:

$$\dot{H}_A = u\left(w_A^H\right) - u\left(w_B^H\right). \tag{8.4}$$

8.2.1 Migrants as Labor Suppliers

We suppose here that the migration of the skilled workers has no significant impact on residents' well-being perhaps because a and b are very small.

Because $H_B = H - H_A$, we have $\dot{H}_B = -\dot{H}_A$ so that we can restrict ourselves to studying the dynamics of H_A. Substituting (8.1) into (8.4) yields the following equation of motion:

$$\begin{aligned}\dot{H}_A &= u[E(H_A)f'(H_A)] - u[E(H - H_A)f'(H - H_A)] \\ &\equiv \phi(H_A).\end{aligned} \tag{8.5}$$

As seen in the foregoing, two opposite effects at work. The equilibrium distribution of skilled workers is the outcome of this trade-off. Because $f'(0) = \infty$ and u is monotonically increasing, it follows that $\dot{H}_A > 0$ when $H_A \to 0$ and $\dot{H}_A < 0$ when $H_A \to H$. Furthermore, the function ϕ defined in (8.5) is continuous in H_A. The theorem of intermediate values therefore implies that at least one equilibrium distribution of the skilled exists at which $\dot{H}_A = 0$ and $0 < H_A < 1$. The preceding boundary conditions mean that all equilibria are interior. In other words, the immobility of the unskilled and the asymptotic property of the production function f imply that all the skilled never concentrate into a single region.

A sufficient condition for a unique, and therefore stable, equilibrium is that ϕ be monotone decreasing. This is so when the function $E(x)f'(x)$ is strictly decreasing in x because u is strictly increasing. This condition means that an increase in the number of skilled in a given region is such that the positive effect of the production externality is always dominated by the negative effect of the marginal productivity decrease, or

$$\frac{E'(x)}{E(x)} < \frac{-f''(x)}{f'(x)} \quad x > 0.$$

This is clearly so when there is no production externality, as in the standard neoclassical model. Such a situation is depicted in Figure 8.1, where the equilibrium distribution of the skilled is symmetric and globally stable.

When there are multiple equilibria, not all equilibria are stable. Accordingly, the emerging equilibrium depends on the initial distribution of skilled workers for the path followed by the economy is unique. To illustrate, assume that the production externality is given by

$$E(H_r) = \exp(\varepsilon H_r) \quad r = A, B, \tag{8.6}$$

where ε is a positive constant. The production function is of the Cobb–Douglas type:

$$F(H_r, L_r) = H_r^{\alpha} L_r^{1-\alpha} \quad \text{or} \quad f(H_r) = H_r^{\alpha} \quad r = A, B \tag{8.7}$$

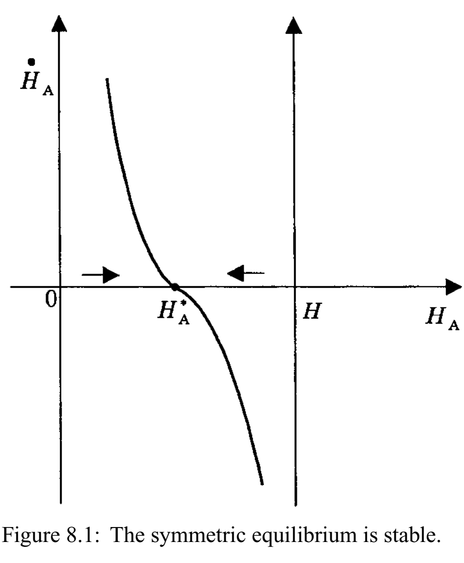


Figure 8.1: The symmetric equilibrium is stable.

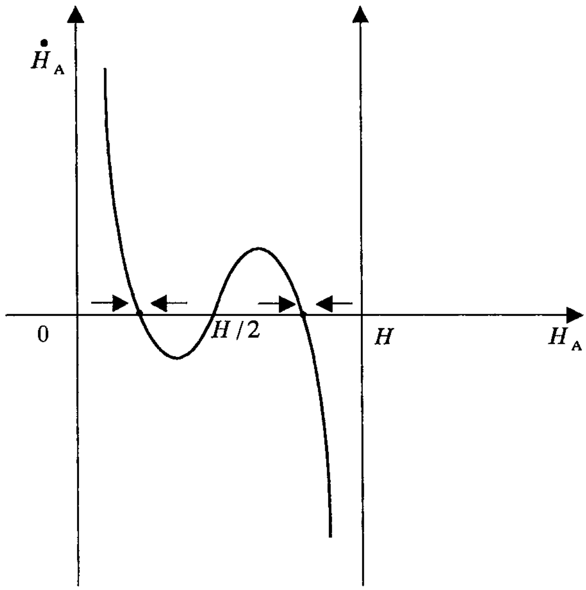


Figure 8.2: The partially agglomerated equilibria are stable.

with $0 < \alpha < 1$, whereas the utility u is

$$u\left(w_r^H\right) = \log\left(w_r^H\right) \quad r = A, B. \tag{8.8}$$

It is shown in Part A of the chapter appendix that there *three* interior equilibria exist (see Figure 8.2) if and only if the following condition holds:

$$\varepsilon H > 2(1 - \alpha), \tag{8.9}$$

meaning that the intensity of the production externality is sufficiently strong compared with the share of the unskilled in the production process.

In this case, the symmetric equilibrium is unstable, whereas the asymmetric equilibria are stable. An asymmetric equilibrium involves more agglomeration when the production externality is stronger, when the cost share of the unskilled is weaker, or both. The region with the larger share of skilled workers will attract more and more workers and will grow at the expense of the other region, which loses a fraction of its skilled workers even when the two regions are initially almost similar. This process shows how a small initial advantage in one region is magnified through the action of the production externality. Although the skilled always earn the same wage in both regions, this is no longer true for the unskilled at each peripheral region: the larger the number of skilled in a region, the higher the wage of the unskilled. Thus, *the economy displays a core–periphery structure when the strength of the production externality is sufficiently strong.* In this case, there is a stable equilibrium in which regions

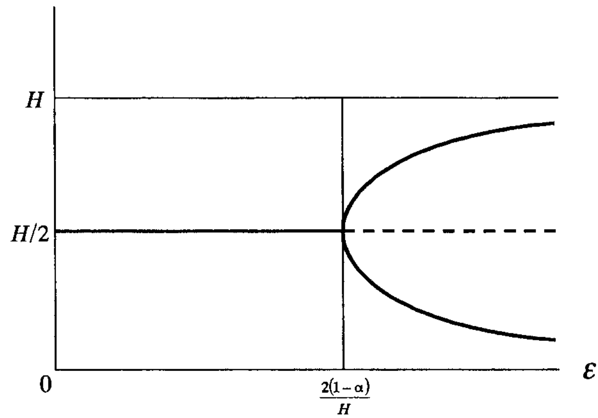


Figure 8.3: Bifurcation diagram of equilibria.

are ex post dissimilar although they are ex ante almost similar. It is worth noting that the reasons for preventing all the skilled from concentrating into a single region are the immobility and the "essentiality" of the unskilled (in the sense of the Inada limit condition). The structural changes in the spatial equilibrium in terms of the production externality parameter ε are depicted in Figure 8.3. Furthermore, although the unskilled initially earn (more or less) the same wage, those residing in the core region end up with higher wages, and those living in the periphery with lower wages, than had been the case initially. Hence, we have here a neat example in which agglomeration generates inequalities within the same group of individuals.

On the contrary, when ε, α, or both are sufficiently small for (8.9) not to hold, the two peripheral equilibria disappear, and there is a single equilibrium involving dispersion. Therefore, at $\varepsilon H = 2(1 - \alpha)$ there is a discontinuity in the set of equilibria.

The preceding discussion is summarized in the following proposition.

Proposition 8.1 *Assume two regions with the same endowment of unskilled who are immobile and essential, whereas regional amenities are independent of the population size. If the production externality is not too strong, a unique stable equilibrium exists, and this equilibrium involves equal dispersion of the skilled workers. However, if the production externality is sufficiently strong, any stable equilibrium involves more than half of the skilled workers in one region, and the wage of the unskilled in the region having the larger number of skilled workers is higher.*

8.2.2 Migrants as Labor Suppliers and Amenity Consumers

We now assume that the migration of the skilled affects the level of amenities in the two regions. Accordingly, the (indirect) utility that an individual obtains in a given region depends not only on the regional wage but also on the various consumption externalities generated in this region by the local population.

Substituting (8.1) into (8.3), we get

$$\dot{H}_A = [u(H_A) + e_A(H_A + 1)] - [u(H - H_A) + e(H - H_A + 1)] \\ \equiv \psi(H_A), \tag{8.10}$$

where $u(x) \equiv u[E(x) f'(x)]$. As in the foregoing, $\dot{H}_A > 0$ when $H_A \to 0$ and $\dot{H}_A < 0$ when $H_A \to H$. Furthermore, ψ is continuous on $[0, H]$. Hence, at least one equilibrium exists, all equilibria are interior, and the equilibrium is globally stable when unique.

Because we consider the case of regions with identical endowments, it can readily be verified that $H_A^* = H_B^* = H/2$ is always an equilibrium. However, this equilibrium may be unstable. To study its behavior, we linearize (8.10) in the neighborhood of $H/2$ and obtain

$$\dot{H}_A \simeq 2m(H_A - H/2)\,[u'(H/2) + e'(H/2 + 1)], \tag{8.11}$$

where m is a positive constant. It then follows from (8.11) that the symmetric equilibrium is stable if and only if

$$u'(H/2) + e'(H/2 + 1) < 0. \tag{8.12}$$

This means that the marginal effect of increasing the number of skilled in any region is negative at the symmetric equilibrium. More precisely, $u'(H/2) < 0$ suggests that the equilibrium wage of the skilled decreases with their number when these workers are equally distributed between the two regions, whereas $e'(H/2 + 1) < 0$ implies that the marginal crowding effect dominates the conviviality effect when the population is evenly split. As before, regional wages are equal inside each group of workers for the symmetrical equilibrium.

When the opposite inequality holds in (8.12), the symmetric equilibrium is unstable, thus implying that the market equilibrium involves some concentration of the skilled even though both regions are a priori identical. Either the production externality or the conviviality effect destroys the stability of the symmetric equilibrium and leads to outcomes in which regions are ex post different. In particular, at any stable asymmetric equilibrium, each type of worker earns different wages in different regions. In other words, persistent wage differentials may well occur even when regions are ex ante identical. For the skilled, this result occurs because earning differential compensates for amenities and congestion differences. This is compatible with an equilibrium state in which the skilled have the same utility level. However, for the unskilled, wage difference reflects differences in marginal productivities as a result of the imbalance in the spatial distribution of skilled workers.

To get a richer characterization of the equilibrium, we return to the special case given by (8.6)–(8.8) as well as by (8.2). It is shown in Part A of the chapter appendix that there are at most three equilibria. The stability properties of these equilibria are similar to those discussed in Section 8.2.1, and equilibrium configurations are qualitatively comparable to those depicted in Figures 8.1, 8.2, and 8.3. Because both regions are ex ante identical, three equilibria exist if and only if

$$\varepsilon + a - b(H+2)/S^2 > 2(1-\alpha)/H. \tag{8.13}$$

This condition reduces to (8.9) when there is no consumption externality ($a = b = 0$). Clearly, inequality (8.13) is likely to hold when the conviviality effect is strong relative to the crowding effect, the production externality is strong, or both. Note that we may have three equilibria even though there is no production externality ($\varepsilon = 0$).

In equilibrium, we must have

$$\{a - b[(H+1)/S)^2\}\,(2H_A - H) = \log\, w_A^H - \log\, w_B^H.$$

Because the derivative of this expression with respect to H_A is equal to

$$2\{a - b[(H+1)/S)^2\} = e'[(H+1)/2]$$

the wage differential varies monotonically with H_A. Consequently, when there are three equilibria, the two stable asymmetric equilibria involve wage dispersion for both types of workers.

We may summarize this discussion as follows:

Proposition 8.2 *Assume two regions with the same endowment of unskilled who are immobile and essential. When the production externality is sufficiently weak, the crowding effect is strong enough, or both, there is a single stable equilibrium that involves full dispersion of the skilled workers. Otherwise, there are several stable equilibria such that more than one half of the skilled are in one region and the wage within each group also varies between regions.*

A few general principles emerge from this analysis.

1. Technological externalities may be sufficient to generate agglomeration and a core–periphery structure even when the economy is perfectly competitive.
2. The geographical immobility of one group of workers (or of one production factor such as land) may be sufficient to preclude the concentration of production in a small number of regions. Indeed, if this group of workers is essential for the production process, the law of diminishing returns is applicable to the other groups, thus reducing the benefit these groups may have from being fully agglomerated. This result fosters the dispersion of production. It is only when the agglomeration externalities are sufficiently large that substantial regional discrepancies may arise.

3. When amenities vary with the population size, the equality of wages between regions does not necessarily hold in equilibrium. Labor mobility aims at equilibrating the combination of wages and external effects.
4. The existence of a production externality implies that the global product is maximized when both groups of workers are concentrated in a single region. As expected, the immobility of the unskilled leads to efficiency losses – at least when amenities are not affected negatively by population size. However, global product as a social objective may be rather a poor goal because narrowing regional discrepancies in per capita incomes may also be a policy objective (Michel, Pestieau, and Thisse 1983). In this perspective, the "emptiness" of a region would appear as a negative outcome.
5. Even when both skilled and unskilled workers are mobile, agglomeration externalities and crowding operate in opposite directions. Thus, the market equilibrium crucially depends on the relative strengths of these two effects, and making predictions about the spatial equilibrium therefore appears to be hard.[8]

8.3 OLIGOPOLY, LOCALIZATION ECONOMIES, AND REGIONAL ADVANTAGE

Consider an industry with two firms (say 1 and 2) each producing a differentiated variety of the same product. Both firms decide first to locate in either of two possible regions (say A and B) and then compete in prices. In order to focus on the pure impact of localization economies, we assume here and in the next section that regions A and B are characterized by the same market conditions. This implies that consumers stay put and do not change their locations with those of the firms considered. More precisely, in each region the demand functions for the two varieties are generated by a representative consumer who has the following quadratic utility function:

$$U(q_1, q_2) = \alpha(q_1 + q_2) - (\beta/2)\left(q_1^2 + q_2^2\right) - \delta q_1 q_2 + z, \tag{8.14}$$

where q_i ($i = 1, 2$) is the quantity of variety i, and z is the quantity of numéraire consumed. As usual, we have $\alpha > 0$ and $0 \leq \delta < \beta$. The consumer's budget constraint is $y = p_1 q_1 + p_2 q_2 + z$.

Maximizing (8.14) subject to the budget constraint yields the standard linear inverse demand schedule $p_i = \alpha - \beta q_i - \delta q_j$ in the price domain where quantities are positive. For $\delta \neq \beta$, the demand function for variety i is given by

$$q_i = a - b p_i + d(p_j - p_i), \tag{8.15}$$

where $a \equiv \alpha/(\beta + \delta)$, $b \equiv 1/(\beta + \delta)$, and $d \equiv \delta/[(\beta - \delta)(\beta + \delta)]$.

The demand system (8.15) may then be interpreted as follows. Parameter d is an inverse measure of the degree of product differentiation between varieties:

they are independent when $d = 0$ (i.e., $\delta = 0$) and perfect substitutes when $d \to \infty$ (i.e., $\beta = \delta$). In other words, increasing the degree of product differentiation between varieties amounts to decreasing d.[9] The own-price effect is stronger (as measured by $b + d$) than each cross-price effect (as measured by d). Finally, parameter b gives the link between individual and industry demand (total demand becomes perfectly inelastic when $b \to 0$, as in the Hotelling model).

To export its product, each firm has to incur a constant unit transportation cost from one region to the other; this cost is given by t units of the numéraire. The production cost structure of the firms depends on their proximity and is described by the following set of assumptions:

- When firms are located in different regions, their marginal cost of production is equal to a constant $c > 0$.
- When firms are located in the same region, they benefit from some positive localization economies. This means that their marginal cost is reduced by a positive amount that may be region-specific. More precisely, if both firms locate in region $r = A, B$, firm i's cost is given by $c - \theta_r$.

In other words, firms experience the same reduction in their marginal production cost when they locate together. However, this reduction may depend on the region where they locate because the nature and intensity of nonmarket interactions between firms may vary from one region to the other. For example, the stock of skills accumulated in a region cannot be moved to another one. A worker operating in a repeated production context continuously uncovers the many facets of the available techniques and gradually adjusts her behavior so as to improve productivity over time. However, workers are heterogeneous not only because they have different skills (by nature as well as by nurture) but also because they face different experiences and have different abilities of learning. When they live in the same locale, they can share their knowledge through various types of social interactions which, in turn, increases their productivity when they are combined within firms. Such features are specific to particular regions. Without loss of generality, it is assumed that the cost reduction is larger in region A than in region B: $\theta_B \leq \theta_A < c$.

We solve the game for its subgame-perfect Nash equilibria by backward induction. We start by solving the second stage of the game. Two different subgames are to be considered according to whether the firms are located together or separately. In both cases, it is supposed that markets are segmented, that is, each firm sets a price specific to the market in which its product is sold (see Section 7.3.3). Indeed, there are many good reasons to believe that firms want to use spatial separation to segment their market (Horn and Shy 1996; Thisse and Vives 1988), and empirical work confirms the assumption that international, not to say interregional, markets are still very segmented (see, e.g., McCallum 1995; Head and Mayer 2000; Wolf 2000).

8.3.1 Interregional Price Competition

1. Assume that both firms are located in region r. Let p_{ir} and q_{ir}, respectively, denote the price and quantity of the product sold by firm i in region r. Firm i's problem entails choosing the prices p_{ir} (the home price) and p_{is} (the foreign price) that maximize its profit function defined as follows:

$$\pi_i = [p_{ir} - (c - \theta_r)][a - bp_{ir} + d(p_{jr} - p_{ir})] + [p_{is} - (c - \theta_r) - t][a - bp_{is} + d(p_{js} - p_{is})]. \tag{8.16}$$

A similar expression holds for firm j.

It is well known that this game has a unique Nash price equilibrium. Taking the first-order conditions and solving for the system of four equations in four unknowns, we obtain the following equilibrium prices:

$$p_{ir} = p_{jr} = \frac{a + (b + d)(c - \theta_r)}{2b + d} \equiv p_r^h \tag{8.17}$$

$$p_{is} = p_{js} = \frac{a + (b + d)(c - \theta_r + t)}{2b + d} \equiv p_r^f,$$

where we use the subscript r to refer to the case in which both firms are located in region r and use h and f to denote variables related to the home and foreign markets.

Equilibrium quantities such as the following are easily found:

$$q_{ir} = q_{jr} = \frac{(b + d)[a - b(c - \theta_r)]}{2b + d} \equiv q_r^h \tag{8.18}$$

$$q_{is} = q_{js} = \frac{(b + d)[a - b(c - \theta_r + t)]}{2b + d} \equiv q_r^f.$$

It can readily be checked that $q_r^h = (b + d)(p_r^h - c + \theta_r)$ and $q_r^f = (b + d)(p_r^f - c + \theta_r - t)$. Thus, the equilibrium quantities and markups are positive (meaning that we have an interior solution and that both firms export to region s) provided that[10]

$$t < \frac{a - bc}{b} - \theta_r. \tag{8.19}$$

Plugging (8.17) and (8.18) into (8.16), we obtain each firm's common equilibrium profits when they are located together in region r as follows:

$$\pi_r = \frac{b + d}{(2b + d)^2}\{[a - b(c - \theta_r)]^2 + [a - b(c - \theta_r + t)]^2\}. \tag{8.20}$$

2. Suppose now that firm i is located in region r and firm j in region $s \neq r$. Firm i's profit function is now written as follows:

$$\pi_i = (p_{ir} - c)[a - bp_{ir} + d(p_{jr} - p_{ir})] + (p_{is} - c - t)[a - bp_{is} + d(p_{js} - p_{is})]. \tag{8.21}$$

Again, a similar expression holds for firm j.

Taking the first-order conditions and solving the corresponding system of four equations, we obtain the following equilibrium prices:

$$p_{ir} = p_{js} = \frac{a + (b+d)c}{2b+d} + \frac{(b+d)dt}{(2b+d)(2b+3d)} \equiv p_S^h$$

$$p_{is} = p_{jr} = \frac{a + (b+d)c}{2b+d} + \frac{2(b+d)^2 t}{(2b+d)(2b+3d)} \equiv p_S^f,$$

where the subscript S refers to the case in which the firms are in separate locations.

It can readily be verified that price differentials between domestic and foreign markets are as follows:

$$p_{is} - p_{ir} = p_{jr} - p_{js} = \frac{b+d}{2b+3d} t < t.$$

Consequently, when firms are separate there is *intraindustry trade* as well as *reciprocal dumping* (Brander and Krugman 1983; Anderson, Schmitt, and Thisse 1995), whereas arbitrage is never profitable.

Equilibrium quantities are then easily computed as follows:

$$q_{ir} = q_{js} = \frac{(b+d)(a-bc)}{2b+d} + \frac{(b+d)^2 dt}{(2b+d)(2b+3d)} \equiv q_S^h$$

$$q_{is} = q_{jr} = \frac{(b+d)(a-bc)}{2b+d} - \frac{(b+d)(2b^2 + 4bd + d^2)t}{(2b+d)(2b+3d)} \equiv q_S^f.$$

As above, one can check that $q_S^h = (b+d)(p_S^h - c)$ and $q_S^f = (b+d)(p_S^f - c - t)$. Using these findings, we can express the condition for an interior solution as

$$t < t_{\text{trade}} \equiv \frac{(2b+3d)(a-bc)}{2b^2 + 4bd + d^2}. \tag{8.22}$$

In what follows, we assume that condition (8.22) is met. Moreover, it is easy to show that condition (8.22) is more stringent than (8.19) for $r = A, B$. In other words, we assume that the transport cost t is low enough to allow firms to export

Table 8.1. First-stage game

	A	B
A	Π_A, Π_A	Π_S, Π_S
B	Π_S, Π_S	Π_B, Π_B

their product whatever their location. Note that condition (8.22) becomes less stringent as products become more differentiated (i.e., as d decreases).

Collecting previous results, we derive the equilibrium profits when the firms locate separately as

$$\pi_S = \frac{b+d}{(2b+d)^2}\left\{\left[a - bc + \frac{(b+d)dt}{2b+3d}\right]^2 + \left[a - bc - \frac{(2b^2+4bd+d^2)t}{2b+3d}\right]^2\right\}. \quad (8.23)$$

8.3.2 Spatial Equilibrium

In the first stage, firms 1 and 2 simultaneously choose their location. Thus, they play the game depicted in Table 8.1.

Comparing expressions (8.21) and (8.23), one can readily verify that $\pi_r > \pi_S$ if and only if $\theta_r > \theta_P(t)$, where

$$\theta_P(t) \equiv \frac{t}{2} - \left(\frac{a-bc}{b}\right) + \sqrt{\frac{(a-bc)(a-bc-bt)}{b^2} + \frac{(b+2d)^2(2b+d)^2}{4b^2(2b+3d)^2}t^2}.$$

It can readily be verified that $\theta_P(t)$ is an increasing function of t and d.

When $\theta_A > \theta_B$, three cases may arise as follows:

Proposition 8.3 *Assume that (8.22) holds. For any triple (θ_A, θ_B, t) such that $\theta_B < \theta_A$, the outcome of the game takes one of the following forms:*

1. *If $\theta_A < \theta_P(t)$, then $\pi_S > \pi_A > \pi_B$, and the equilibrium is unique (up to a permutation) and involves dispersion.*
2. *If $\theta_B < \theta_P(t) < \theta_A$, then $\pi_A > \pi_S > \pi_B$ and the unique equilibrium involves agglomeration in region A.*
3. *If $\theta_P(t) < \theta_B < \theta_A$, then $\pi_A > \pi_B > \pi_S$ and there are two equilibria involving agglomeration in region A or in region B (the former equilibrium Pareto-dominates the latter).*[11]

In words, for a given value of the transportation cost, this means that firms must be compensated by a sufficient amount of localization economies for the increased competition resulting from a common location. Or, to put it differently, *when transport costs are low, agglomeration is the market outcome because each firm can benefit from a production cost reduction by being together without losing much business in the other region.* It is worth noting that the threshold level $\theta_P(t)$ decreases when the degree of product differentiation rises (that is, when d falls). Indeed, more product differentiation relaxes price competition and makes the agglomeration of firms more likely. On the other hand, $\theta_P(t)$ rises with the transportation cost because higher trade costs strengthen the benefits of geographical isolation. If there were no localization economies ($\theta_A = \theta_B = 0$), equilibrium would always involve dispersion.

The preceding inequalities may be reinterpreted in the context in which firm 1 is already located when firm 2 considers entering the market. If $\theta_P(t) < \theta_B < \theta_A$, firm 2 always wants to be with firm 1 regardless of its location. Hence, if firm 1 is in region B for some historical reasons, the agglomeration will occur there even though this region is less efficient. This result provides another illustration of the phenomenon of *lock-in effect* associated with the presence of Marshallian externalities and shows once more how history matters in the development of a particular region.[12]

8.3.3 Welfare

Plugging the linear inverse demands into function (8.14) and using the definitions of b and d, we can readily derive the general formulation of the consumers' surplus as follows:

$$C = \frac{1}{2b(b+2d)}\left[(b+d)\left(q_1^2 + q_2^2\right) + 2dq_1q_2\right]. \tag{8.24}$$

We adopt the following notation. Let C_{s} denote the surplus for the consumers in region $\mathrm{s} = A, B$, and let the location of the firms be represented by r if both firms are in $r = A, B$, or S if they are in separate locations. Similarly, let $C(r) = C_A(r) + C_B(r)$ and $C(\mathrm{S}) = C_A(\mathrm{S}) + C_B(\mathrm{S})$, respectively, denote the total consumer surplus when both firms are in region r or when they are separate. Then, (8.24) yields

$$C_r(r) = (1/b)\left(q_r^h\right)^2 \qquad r = A, B \quad \text{and} \quad C_r(\mathrm{s}) = (1/\mathrm{b})\left(\mathrm{q}_{\mathrm{s}}^{\mathrm{f}}\right)^2 \qquad r \neq \mathrm{s}$$

$$C_r(\mathrm{S}) = \frac{1}{2b(b+2d)}\left\{(b+d)\left[\left(q_{\mathrm{S}}^h\right)^2 + \left(q_{\mathrm{S}}^f\right)^2\right] + 2dq_{\mathrm{S}}^h q_{\mathrm{S}}^f\right\} \equiv C_{\mathrm{S}}, \quad r = A, B.$$

We start by considering the *first best* situation in which the planner is able to control both the locations of firms and their prices, whereas lump-sum transfers are available. Because of marginal cost pricing, the following quantities are supplied:

$$q_r^h = a - b(c - \theta_r) \quad \text{and} \quad q_r^f = a - b(c - \theta_r + t) \qquad r = A, B$$
$$q_S^h = a - bc \quad \text{and} \quad q_S^f = a - b(c + t).$$

Because firms earn zero profits, the global welfare is equal to the global consumer surplus. A simple calculation reveals that $C(A) \geq C(B) > C(\mathrm{S})$. In other words, we have the following:

Proposition 8.4 *Assume that (8.22) holds. If $\theta_A > \theta_B$, then the first best optimum always involves agglomeration in region A. If $\theta_A = \theta_B$, then the first best optimum involves agglomeration in region A or in region B.*

This proposition implies that it is always socially desirable that firms be agglomerated when trade occurs. Yet, when the intensity of the localization economies in region A is not large ($\theta_P(t) > \theta_A$), strategic competition leads to dispersion. However, the reverse does not hold because the market never yields excessive agglomeration of firms in this setting.

When transport costs are low enough, the market location outcome is likely to coincide with the first best location pattern. In this case, the efficiency loss arises only from the discrepancy between prices and marginal costs. Nevertheless, the market is at the origin of another efficiency loss when agglomeration arises in region B, whereas it is socially desirable that firms located together in region A ($\theta_P(t) < \theta_B < \theta_A$).[13] This happens when one firm first establishes itself in a region that offers a low potential for the development of localized externalities.

Next, we consider a *second best* situation in which the planner is able to control the locations of firms but not their prices and quantities, which are determined at the market equilibrium. Plugging equilibrium quantities into the preceding expressions, we have that the consumers in region r earn the following surpluses according to whether (1) two firms are located in their region, or (2) one or zero firms is located in region r:

$$C_r(r) = \frac{(b+d)^2}{b(2b+d)^2}[a - b(c - \theta_r)]^2$$

$$C_r(s) = \frac{(b+d)^2}{b(2b+d)^2}[a - b(c - \theta_s + t)]^2$$

$$C_\mathrm{S} = \frac{(b+d)^2}{b(2b+d)^2}\left[(a - bc)(a - bc - bt) + \frac{(b+d)(4b^2 + 8bd + d^2)bt^2}{2(2b+3d)^2}\right].$$

Let us first adopt the viewpoint of local governments. From our assumptions that $\theta_A \geq \theta_B$ and $t > \theta_A - \theta_B$, it is easily seen that consumer surpluses in the case of common location are ranked as follows: $C(A) \geq C_B(B) > C_B(A) > C_A(B)$.

Furthermore, if we compare the consumer surpluses when firms are located either together or separately, we can establish the following two results: (1) $C_A(A) \geq C_B(B) > C_S$; (2) $C_r(L) > C_S$ if and only if $\theta_s > \theta_R(t)$, where

$$\theta_R(t) \equiv t - \left(\frac{a-bc}{b}\right) + \sqrt{\frac{(a-bc)(a-bc-bt)}{b^2} + \frac{(b+d)(4b^2+8bd+d^2)}{2b(2b+3d)^2}t^2}.$$

These findings are summarized in the following proposition:

Proposition 8.5 *Assume that $t > \theta_A - \theta_B$. The second best optimum is such that*

1. *consumers in region $r = A, B$ are better off when both firms locate in their region than when only one does so;*
2. *if the intensity of localization economies in region r exceeds some threshold $\theta_R(t)$, even consumers in region $s \neq r$ are better off when no firm locates in their region than when one does.*

Thus, a local government should not necessarily strive to attract a firm when localization economies are strong because such a policy, when it is effective, can make the members of its constituency worse off. Of course, such considerations do not account for the welfare gain associated with the creation of jobs accompanying the establishment of a new firm in this region. This analysis also reveals a possible conflict of interest between workers who would find a job in the new company and the whole body of consumers living in the region.[14]

We now consider the global welfare. As far as firm's interests are concerned, we already know that total profits are higher when both firms locate in region r than when they separate provided that $\theta_r > \theta_P(t)$. Comparing global surpluses for the consumers for different locations, we see that $C(r) > C(\mathrm{S})$ if and only if $\theta_r > \theta_C(t)$, where

$$\theta_C(t) \equiv \frac{t}{2} - \left(\frac{a-bc}{b}\right) + \sqrt{\frac{(a-bc)(a-bc-bt)}{b^2} + \frac{(b+2d)(2b+d)^2}{4b(2b+3d)^2}t^2}.$$

Let the global welfare W be defined as the sum of total profits and total consumer surplus. It can be shown that $W(r) > W(\mathrm{S})$ if and only if $\theta_r > \theta_W(t)$, where

$$\theta_W(t) \equiv \frac{t}{2} - \left(\frac{a-bc}{b}\right) + \sqrt{\frac{(a-bc)(a-bc-bt)}{b^2} + \frac{(3b+5d)(b+2d)(2b+d)^2}{4b(3b+d)(2b+3d)^2}t^2}.$$

Some straightforward computations reveal that $\theta_P(t) > \theta_W(t) > \theta_C(t)$ for all t smaller than t_{trade}. It then appears that the second best outcome may involve agglomeration even though the market selects dispersion, but the reverse is never true.

In addition, the preceding inequalities also mean that the interests of the various economic groups may be vastly different in the choice of a locational pattern for firms. For instance, if for a given value of the transportation cost t we have that $\theta_W(t) < \theta_A < \theta_P(t)$, then firms choose separate locations, whereas the federal government would prefer the firms to be agglomerated in region A at the second best optimum. By contrast, when $\theta_C(t) < \theta_A < \theta_W(t)$, both the federal government and firms prefer separate locations, but consumers as a whole are better off when agglomeration occurs in region A.

More conflict might even appear if the interests of consumers in each region are taken into account. It is indeed possible to have situations in which $\theta_W(t) < \theta_A < \theta_C(t) < \theta_P(t)$. Then, firms choose separate locations in accordance with the interests of consumers in region B but not with the interest of the consumers of region A and of the federal government (which would prefer agglomeration in A).

In summary, the existence of localization economies may explain why oligopolistic firms want to be together despite the fiercer competition triggered by a common location. As in Chapter 7, such a clustering often appears to be globally efficient and may even lead to an increase in the welfare of consumers leaving the region that does not accommodate the cluster. However, this gain in efficiency is obtained at the expense of a higher degree of spatial inequity. In addition, the welfare analysis of the various groups involved reveals the existence potential conflicts of interest, thus suggesting that policy recommendations are not easy to make once it is recognized that lump-sum transfers are not available.

8.4 THE FORMATION OF INDUSTRIAL CLUSTERS UNDER LOCALIZATION ECONOMIES

In this section, we consider an industry with a continuum of a given size M of firms producing each a variety of a (horizontally) differentiated product. The representative consumer's utility function is now as follows:[15]

$$U(z; q(i), i \in [0, M]) = \alpha \int_0^M q(i)di - \frac{\beta - \delta}{2} \int_0^M [q(i)]^2 di - \frac{\delta}{2}\left[\int_0^M q(i)di\right]^2 + z, \tag{8.25}$$

where $q(i)$ is the quantity of variety $i \in [0, M]$ and z the quantity of the numéraire, whereas the parameters in (8.25) are such that $\alpha > 0$ and $\beta > \delta > 0$. In this expression, α is a measure of the intensity of preferences for the differentiated product with respect to the numéraire, whereas $\beta > \delta$ means that the representative consumer has a love for variety. Suppose, indeed, that a consumer consumes a total of $Mq \equiv \int_0^M q(i)di$ of the differentiated product such that the consumption is uniform on $[0, x]$ and zero on $(x, M]$. Then, the density on $[0, x]$ is Mq/x. Equation (8.25) evaluated at this consumption pattern is

$$\begin{aligned} U &= \alpha \int_0^x \frac{Mq}{x} di - \frac{\beta - \delta}{2} \int_0^x \left(\frac{Mq}{x}\right)^2 di \\ &\quad - \frac{\delta}{2} \int_0^x \int_0^x \left(\frac{Mq}{x}\right)^2 di dj + z \\ &= \alpha Mq - \frac{\beta - \delta}{2x} M^2 q^2 - \frac{\delta}{2} M^2 q^2 + z, \end{aligned} \tag{8.26}$$

which is increasing in x and, hence, maximized at $x = M$, where variety consumption is maximal. We may then conclude that, when $\beta > \delta$, the quadratic utility function exhibits a preference for variety.[16] Finally, for a given value of β, the parameter δ expresses the substitutability between varieties: the higher δ is, the closer substitutes are the varieties.

The consumer is endowed with $\bar{z} > 0$ units of the numéraire. Her budget constraint can then be written as follows:

$$\int_0^M p(i)q(i)di + z = \bar{z},$$

where $p(i)$ is the price of variety i. The initial endowment $\bar{z}$ is supposed to be sufficiently large for the optimal consumption of the numéraire to always be strictly positive at the market outcome. Solving the budget constraint for the numéraire consumption, plugging the corresponding expression into (8.25), and taking the first-order conditions with respect to $q(i)$, we obtain

$$\alpha - (\beta - \delta)q(i) - \delta \int_0^M q(j)dj = p(i) \qquad i \in [0, M],$$

which gives the demand function for variety $i \in [0, M]$:[17]

$$\begin{aligned} q(i) &= a - bp(i) + d \int_0^M [p(j) - p(i)]dj \\ &= a - (b + dM)p(i) + dP, \end{aligned} \tag{8.27}$$

where $a \equiv \alpha/[\beta + (M-1)\delta]$, $b \equiv 1/[\beta + (M-1)\delta]$, $d \equiv \delta/(\beta - \delta)[\beta + (M-1)\delta]$, and

$$P = \int_0^M p(j)dj$$

so that P/M may be interpreted as the price index. As in the preceding section, the parameter d represents an inverse measure of the degree of product differentiation among varieties: decreasing d amounts to increasing the degree of product differentiation among the given set of varieties. Furthermore, the own-price effect is stronger (as measured by $b + dM$) than each cross-price effect (as measured by d), thus allowing for different elasticities of substitution between pairs of varieties. Because the mass of firms is fixed and given, we may assume without substantial loss of generality that $M = 1$, and thus $a \equiv \alpha/\beta$, $b \equiv 1/\beta$, and $d \equiv \delta/\beta(\beta - \delta)$.[18] The more general case in which M may differ from 1 will be used in Chapter 9.

As in the preceding section, there are two regions that have the same market conditions. It is also assumed that the two regions are homogeneous with respect to the law governing localization economies: when there are M_r firms in region r, firm i is able to produce the variety i at marginal cost $c_r \equiv c(M_r)$. Of course, we must have $M_A + M_B = 1$. Let t be the unit transport cost between the two regions expressed in terms of the numéraire. It is assumed that

$$t < \frac{a}{b} - \max\{c(M_A), c(M_B)\}, \tag{8.28}$$

meaning that even a firm facing the higher marginal cost always finds it profitable to export to the other region.

8.4.1 Equilibrium Pricing

We study here the process of competition between firms for a given spatial distribution (M_A, M_B) of firms. Because we have a continuum of firms, each one is negligible in the sense that its action has no impact on the market. Hence, when choosing its prices, a firm accurately neglects the impact of its decision over the prices selected by the other firms. In addition, because firms sell differentiated varieties, each firm has some monopoly power because it faces a demand function with finite elasticity. All of this is in accordance with Chamberlin's (1933) large group competition in which the effect of a price change by one firm has a significant impact on its own demand but only a negligible impact on competitors' demands. However, a firm must account for the distribution of the firms' prices through some aggregate statistics, given here by the *price index*, to find its equilibrium price. Consequently, our market solution is given by a Nash equilibrium with a continuum of players in which prices are interdependent: each firm neglects its impact on the market but is aware that the market as a whole has a nonnegligible impact on its behavior.

Again, we assume that firms compete in segmented markets. In what follows, we suppose that the parameters of the economy are such that the equilibrium prices exceed costs and markups are positive (meaning that we have an interior solution and that exportation occurs for all firms). A sufficient condition for this assumption to hold will be given below.

Assume that variety i is produced in $r = A, B$. The corresponding firm sells its output on both markets, that is, quantity $q_{rr}(i)$ at price $p_{rr}(i)$ on market r, and quantity $q_{rs}(i)$ at price $p_{rs}(i)$ on market $s \neq r$. Thus, $p_{rr}(i)$ is the price in region r of variety i produced locally, whereas $p_{rs}(i)$ the price of the same variety exported from r to s. We adopt the following notation:

$$P_{rr} \equiv \int_{i \in M_r} p_{rr}(i)\,di \quad \text{and} \quad P_{rs} \equiv \int_{i \in M_r} p_{rs}(i)\,di \qquad s \neq r.$$

Demands for firm i located in r are then given by

$$q_{rr}(i) = a - (b + d)p_{rr}(i) + d(P_{rr} + P_{sr}) \qquad s \neq r$$

and

$$q_{rs}(i) = a - (b + d)p_{rs}(i) + d(P_{rs} + P_{ss}) \qquad s \neq r.$$

Firm i in r maximizes its profits defined by

$$\pi_r(i) = [p_{rr}(i) - c_r]q_{rr}(i) + [p_{rs}(i) - c_r - t]q_{rs}(i) \qquad s \neq r.$$

Deriving the first-order conditions for profit maximization with respect to prices and summing them for all firms located in r, we obtain the following expressions:

$$[2(b + d) - dM_r]P_{rr} - dM_r P_{sr} = M_r[a + (b + d)c_r] \qquad s \neq r \tag{8.29}$$

$$[2(b + d) - dM_A]P_{rs} - dM_r P_{ss} = M_r[a + (b + d)(c_r + t)] \qquad s \neq r. \tag{8.30}$$

Varieties being symmetric, in equilibrium we have

$$p_{rs}(i) = p_{rs} = P_{rs}/M_r.$$

Because profit functions are concave in own price, solving the system of equations (8.29)–(8.30) yields the equilibrium prices

$$p^*_{rr} = \frac{c_r}{2} + \frac{2a + d[M_r c_r + M_s(c_s + t)]}{2(2b + d)}$$

$$p^*_{rs} = \frac{c_r + t}{2} + \frac{2a + d[M_r(c_r + t) + M_s c_s]}{2(2b + d)} \qquad s \neq r.$$

As expected, the equilibrium prices depend on the distribution of firms between the two regions. Prices rise when the size of the local market, evaluated by a, gets larger or when the degree of product differentiation, inversely measured by d, increases, provided that (8.28) holds. All these results agree with what is known in industrial organization and spatial pricing theory. By inspection, it can also be readily verified that the local price, p^*_{rr}, increases with t because the local firms in r are more protected against distant competitors, whereas the export price, $p_{rs} - t$, decreases because it becomes more difficult for these firms to penetrate the distant market. Finally, both the prices charged by local and distant firms fall when the number (or mass) of local firms in region r increases, while the total number of firms is held constant, if and only if $c_r < c_s + t$. This occurs because the lower cost prevailing in r intensifies local price competition.

Using the first-order conditions, we can easily establish the following relationships between equilibrium prices and quantities:

$$q_{rr} = (b + d)(p_{rr} - c_r) \quad \text{and} \quad q_{rs} = (b + d)(p_{rs} - c_r - t).$$

The equilibrium profits of any firm located in region $r \neq s$ are thus

$$\begin{aligned}\pi_r(M_r, M_s) &= (p_{rr} - c_r)q_{rr} + (p_{rs} - c_r - t)q_{rs} \\ &= (b + d)[(p_{rr} - c_r)^2 + (p_{rs} - c_r - t)^2] \\ &= \frac{b + d}{2(2b + d)^2}\big\{[2(a - bc_r) - dM_s(c_r - c_s) - bt]^2 \\ &\qquad + (b + dM_s)^2t^2\big\}.\end{aligned}$$

In the remainder of the section, it is assumed that localization economies obey the same law in each region $r = A, B$ given by

$$c_r(M_r) = c - \theta M_r$$

where $0 < \theta < c$. We are then able to state the conditions under which the equilibrium prices and quantities are positive (meaning that we have an interior solution and that exportation occurs for all firms). It can readily be checked that a sufficient condition is that $q_{rs} > 0$ in the limiting case in which $M_r = 0$,

$$t < t'_{\text{trade}} \equiv \frac{2(a - bc) - d\theta}{2b + d}, \tag{8.31}$$

whose right-hand side is supposed to be positive. Note that (8.28) is satisfied when (8.31) holds.

8.4.2 Spatial Equilibrium

We can take advantage of the symmetry of the problem by setting $\Delta M = M_A - M_B$. Thus, $M_A = (1/2)(1 + \Delta M)$, $M_B = (1/2)(1 - \Delta M)$, $c_A(M_A) + c_B(M_B) = 2c - \theta$, and $c_A(M_A) - c_B(M_B) = -\theta\Delta M$. Consequently, the equilibrium profits can be rewritten as follows (where D stands for $2b + d$):

$$\pi_A(\Delta M) = \frac{b+d}{8D^2}\{[4a - 2b(2c + t - \theta) - d\theta(\Delta M)^2 + D\theta(\Delta M)]^2 + [D - d(\Delta M)]^2 t^2\}$$

$$\pi_B(\Delta M) = \frac{b+d}{8D^2}\{[4a - 2b(2c + t - \theta) - d\theta(\Delta M)^2 - D\theta\Delta M]^2 + [D + d(\Delta M)]^2 t^2\}.$$

Accordingly, the difference $\Delta\pi(\Delta M)$ between the profits earned in each region is given by

$$\Delta\pi(\Delta M) = \frac{b+d}{2D}\Delta M\{[4a - 2b(2c + t - \theta) - d\theta(\Delta M)^2]\theta - dt^2\},$$

which can be rewritten as the following cubic function of ΔM:

$$\Delta\pi(\Delta M) = \frac{d\theta^2(b+d)}{2D}\Delta M[\Lambda - (\Delta M)^2], \tag{8.32}$$

where Λ is a constant given by

$$\Lambda \equiv \frac{[4a - 2b(2c + t - \theta)]\theta - dt^2}{d\theta^2}, \tag{8.33}$$

whereas its slope is

$$\frac{d(\Delta\pi)}{d(\Delta M)} = -\frac{d\theta^2(b+d)}{2D}[3(\Delta M)^2 - \Lambda]. \tag{8.34}$$

The distribution $M_A \in [0, 1]$ is a *spatial equilibrium* when no firm may earn a higher profit by changing location. This arises at an interior point $M_A \in (0, 1)$ when

$$\Delta\pi(2M_A - 1) = 0$$

or at $M_A = 0$ when $\Delta\pi(-1) \leq 0$, or at $M_A = 1$ when $\Delta\pi(1) \geq 0$. In the first case, we have either a *fully dispersed configuration* or a *partially agglomerated configuration*; in the last two cases, we have a *fully agglomerated configuration.*

Given (8.32), the symmetric configuration ($\Delta M = 0$) is always an equilibrium. However, because other spatial equilibria exist, it is convenient to use stability as a selection device because an unstable equilibrium is unlikely to happen. To study the stability of a spatial equilibrium, we must define how firms behave away from the equilibrium. In the present context, the dynamics is fairly natural. If firms observe that a region offers higher profits than the other, they want to move to that location. In other words, the driving force in the *profit* *differential* between A and B,

$$\dot{M}_A = M_A \Delta\pi(\Delta M) M_B,$$

where $\dot{M}_A$ is the time-derivative of M_A. If $\Delta\pi$ is positive and if $0 < M_A < 1$, some firms will move from B to A; if it is negative, some will go in opposite direction. An equilibrium is *stable* if, for any marginal deviation from the equilibrium, the equation of motion above brings the distribution of firms back to the original one. Therefore, a fully agglomerated configuration is always stable when it turns out to be an equilibrium, whereas an interior equilibrium is stable if and only if the slope of $\Delta\pi(\Delta M)$ is nonpositive in a neighborhood of this equilibrium. Clearly, a spatial equilibrium is a steady-state for the equation $\dot{M}_A = 0$, and conversely.

Several kinds of stable equilibria may arise in the present setting. Either all firms agglomerate in one region (corner solution) or they distribute themselves between the two regions (interior solution) in a way that equalizes profits. In the latter case, firms can spread evenly ($\Delta M = 0$) or unevenly across regions. These equilibria are now fully characterized.

Proposition 8.6 *The two-region economy always has a single (up to a permutation) stable spatial equilibrium. This one involves the following:*

1. *symmetry* ($\Delta M = 0$) *if and only if* $\Lambda \leq 0$;
2. *partial agglomeration* ($\Delta M = \pm\sqrt{\Lambda}$) *if and only if* $0 < \Lambda < 1$;
3. *full agglomeration* ($\Delta M = \pm 1$) *if and only if* $1 \leq \Lambda$.

Proof On the basis of (8.32), the proof involves three steps.

1. When $\Lambda < 0$, the equation $\Delta\pi = 0$ has a single real solution $\Delta M = 0$ at which the slope is negative. Thus, $\Delta M = 0$ is the unique equilibrium and is stable.
2. When $\Lambda \geq 0$, the equation $\Delta\pi = 0$ has three real solutions ($\Delta M = 0$ and $\Delta M = \pm\sqrt{\Lambda}$) and one of the following two cases must arise:
   A. The asymmetric solutions are equilibria if and only if $0 < \Lambda < 1$. They are stable because the slope of $\Delta\pi$ evaluated at $\Delta M = \pm\sqrt{\Lambda}$ is nonpositive as $0 < \Lambda < 1$.
   B. When $\Lambda > 1$, there is no asymmetric interior equilibrium, and the only stable equilibria are such that $\Delta M = \pm 1$ because either $\Delta\pi(1) > 0$ or $\Delta\pi(-1) > 0$. Q.E.D.

The stable equilibria are represented by the bold lines and the unstable ones by the dotted lines in Figure 8.4.

Despite the symmetry of the economy, *industrial clusters involving different numbers of firms may emerge in equilibrium.* In this case, the number of

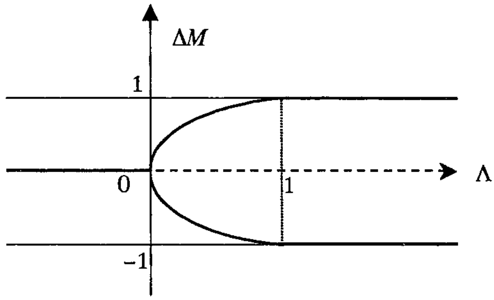


Figure 8.4: Stable equilibria.

firms in region r is

$$M_r^* = \frac{1 \pm \sqrt{\Lambda}}{2}.$$

Of course, the region ending up with the larger number of firms is the one that has the larger initial share of firms regardless of how small the difference is. This shows again that history (here the initial conditions) matters for the geographical distribution of production. In other words, the existence of localization economies may lead to the emergence of a polarized space, especially when transport costs are low.

Because ΔM depends only upon Λ, the impact of each parameter on the spatial equilibrium can be analyzed through Λ given by (8.33). First, as the intensity of preference for the output of the industry (a) rises, the degree of asymmetry between the two clusters grows. Second, we have

$$\frac{\partial \Lambda}{\partial t} = -\frac{2(b\theta + dt)}{d\theta^2} < 0 \tag{8.35}$$

$$\frac{\partial \Lambda}{\partial d} = -\frac{\Lambda}{d} - \frac{t^2}{d\theta^2} < 0 \tag{8.36}$$

$$\frac{\partial \Lambda}{\partial \theta} = -\frac{d\theta^2\Lambda + 2\theta^2 b - t^2 d}{d\theta^3} = -\frac{\Lambda}{\theta} + 2\frac{b}{d\theta} + \frac{t^2}{\theta^3}. \tag{8.37}$$

Equation (8.35) shows that a decrease in transport cost leads to more asymmetry between clusters. This finding reveals that lower transportation costs are likely to drive the economy toward more agglomeration in one region at the expense of the other. It is worth pointing out that this process is smooth unlike what is observed in the core–periphery model we will study in the next chapter. Similarly, as shown by (8.36), more product differentiation leads to more agglomeration of firms within the large region. As seen in Chapter 7, this is a standard result in many spatial competition models.

Equation (8.37) shows that an increase in the intensity of localization economies strengthens the tendency toward agglomeration provided that θ is small enough, that is,

$$\theta < \frac{2dt^2}{4(a - bc) - bt}, \tag{8.38}$$

where the right-hand side is positive by (8.28). In contrast, when (8.38) does not hold, we see that stronger localization economies generate more dispersion. This surprising result may be explained as follows. When production costs are not too high, firms in the small cluster price in the inelastic part of their demands, whereas firms in the large cluster price in the elastic part. By reducing production costs, an increase in θ thus intensifies competition much more in the large cluster than in the small one, leading some firms to move from the large to the small cluster. In sum, localization economies do not have a monotone impact on the firm distribution.

8.4.3 Welfare

The research strategy is similar to the one used above. To begin with, we assume that the planner is able to impose prices equal to marginal costs as well as to choose firms' locations. Because firms earn zero profits, the only relevant welfare measure is the global consumer surplus $C_{\mathrm{G}} = \sum_{s=A,B} C_s$. It can readily be verified that the consumer surplus is given by the following expression:

$$\begin{aligned} C &= \alpha \int_0^1 q(i)di - \frac{1}{2}(\beta - \delta) \int_0^1 [q(i)]^2 di \\ &\quad - \frac{1}{2}\delta\left(\int_0^1 q(i)di\right)^2 - \int_0^1 p(i)q(i)di \\ &= \frac{1}{2}(\beta - \delta)\int_0^1 [q(i)]^2 di + \frac{1}{2}\delta\left(\int_0^1 q(i)di\right)^2 \\ &= \frac{1}{2b(b+d)}\left[b\int_0^1 [q(i)]^2 di + d\left(\int_0^1 q(i)di\right)^2\right], \end{aligned}$$

and thus

$$\begin{aligned} C_s &= \frac{1}{2(b+d)}\left[(q_{ss}^*)^2 N_s + (q_{rs}^*)^2 N_r\right] \\ &\quad + \frac{d}{2b(b+d)}\left[N_s^2(q_{ss}^*)^2 + N_r^2(q_{rs}^*)^2 + 2(N_s q_{ss}^*)(N_r q_{rs}^*)\right], \end{aligned} \tag{8.39}$$

where the first subscript denotes the region in which consumption occurs and the second the region in which the varieties are produced ($r, s = A, B$ and $r \neq s$).

After having replaced prices and quantities by their first best values, we obtain

$$C_{\mathrm{G}} = -\frac{A}{4}(\Delta M)^4 + \frac{A}{2}\Gamma(\Delta M)^2 + B,$$

where

$$A \equiv d\theta^2 > 0$$

and B are constants and

$$\Gamma \equiv \frac{(3b+d)\theta^2 + 4\theta(a - bc - bt/2) - dt^2}{2d\theta^2} = \frac{b+d}{2d} + \frac{\Lambda}{2}. \tag{8.40}$$

The first best locational pattern is then obtained by maximizing C_{G} with respect to ΔM.

Proposition 8.7 *The first best allocation involves*

1. *full dispersion if and only if* $\Gamma \leq 0$*;*
2. *partial agglomeration if and only if* $0 < \Gamma < 1$*;*
3. *full agglomeration if and only if* $1 \leq \Gamma$.

Proof

1. We seek $\Delta M \in [-1, 1]$, maximizing C_{G}. When the maximum is interior, we have

$$\frac{\partial C_{\mathrm{G}}}{\partial(\Delta M)} = -A\Delta M[(\Delta M)^2 - \Gamma] = 0 \tag{8.41}$$

and

$$\frac{\partial^2 C_{\mathrm{G}}}{\partial(\Delta M)^2} = -A[3(\Delta M)^2 - \Gamma] < 0.$$

When $\Gamma \leq 0$, there is a unique maximizer given by $\Delta M = 0$ because the second-order condition is satisfied.

2. When $0 < \Gamma < 1$, C_{G} reaches a minimum at $\Delta M = 0$ and is maximized at $\Delta M = \pm\sqrt{\Gamma}$, where the second-order condition holds.
3. Finally, when $1 \geq \Gamma$, the welfare function C_{G} is maximized at $\Delta M = \pm 1$. Q.E.D.

Thus, the first best solution displays a pattern similar to that arising when firms are free to choose prices and locations at the market equilibrium. In the case of asymmetric clusters, the socially optimal distribution of firms (M_A^0, M_B^0) is given by

$$M_A^0 = \frac{1 \pm \sqrt{\Gamma}}{2}.$$

It is worth noting that Γ is a strictly decreasing function of t and d as shown by differentiating (8.40) with respect to t and d,

$$\frac{\partial \Gamma}{\partial t} = \frac{1}{2}\frac{\partial \Lambda}{\partial t} < 0 \tag{8.42}$$

$$\frac{\partial \Gamma}{\partial d} = -\frac{b}{2d^2} + \frac{1}{2}\frac{\partial \Lambda}{\partial d} < 0, \tag{8.43}$$

and thus *lower transport costs, more product differentiation, or both yield more asymmetry between optimal clusters.* From (8.40), it can readily be verified that the impact of a or θ on Γ is similar to the impact on Λ, as discussed in Section 8.4.2.

Furthermore, it follows from (8.40) that $\Gamma > 0$ when $\Lambda = 0$ and that $\Gamma = \Lambda$ for a value of Λ that exceeds 1. Consequently, we have the following result:

Proposition 8.8 *The first best optimum never involves a lower degree of agglomeration than the market equilibrium.*

Thus, as in the duopoly case, we observe that, with a large group of firms, the large cluster never involves too many firms in equilibrium from the efficiency viewpoint. In particular, the planner sets up more asymmetric clusters than arise at the market solution unless the market outcome corresponds to full agglomeration. This requires some explanation. At the optimum, prices are set at the marginal cost level whereas locations are chosen to maximize the difference between the benefits of agglomeration and total transport costs. By contrast, at the market equilibrium, firms take advantage of their spatial separation to relax price competition and, hence, to make higher profits. These two effects combine to generate the discrepancy between the market and optimal solutions. This confirms what we saw in Section 7.3, that is, price competition is a strongly dispersion force.

Unless full dispersion corresponds to both the equilibrium and the optimum, the difference between regional surpluses generates a conflict between regions about firms' locations. Indeed, the region with the larger cluster benefits from larger localization economies, and thus lower prices, as well as from lower transportation costs on its imports (through less varieties and smaller quantities). This occurs because the planner focuses only on global efficiency and not on interregional equity. This policy is a sensible one when lump-sum transfers compensating the consumers of the less industrialized region are available. However, when such redistributive instruments are not available, a trade-off between global efficiency and interregional equity arises.

Consider now a situation in which the planner is able to control firms' locations but not their prices, which are determined at the market equilibrium. Assume again that the planner's purpose is to maximize the global surplus. The consumer surplus C_s in region $s = A, B$ is computed by plugging the market equilibrium quantities into expression (8.39). The producer surplus Π_r in region $r = A, B$ is given by the sum of profits made in r and s by the firms located in r:

$$\Pi_r = \frac{M_r}{b+d}\left[(q^*_{rr})^2 + (q^*_{rs})^2\right].$$

To determine the second best, it is therefore sufficient to maximize the following welfare function:

$$W_{\mathrm{G}} = C_A + C_B + \Pi_A + \Pi_B.$$

The first-order condition yields a cubic function with properties similar to those obtained in the first best when Γ is replaced by Θ as given by

$$\Theta \equiv \frac{4(3b+d)\theta[4a - 2b(2c+t-\theta)] + 4b(3b+d)\theta^2 - d(8b+3d)\,(t^2-\theta^2)}{2d\theta^2(8b+3d)}$$

$$= \frac{3(2b+d)^2 + d(4b+d)\,(t/\theta)^2}{2d(8b+3d)} + \frac{2(3b+d)}{8b+3d}\Lambda.$$

Consequently, the second-best allocation involves (1) identical clusters if and only if $\Theta \leq 0$; (2) asymmetric clusters if and only if $0 < \Theta < 1$; (3) a single cluster if and only if $1 \leq \Theta$.

The comparison between the second best and the market outcomes is very similar to that made above provided that Λ and Θ move in the same direction. In particular, it can readily be verified that Proposition 8.7 still holds, that is, $\Theta > \Lambda$ because $\Theta > 0$ when $\Lambda = 0$, whereas $\Lambda = \Theta$ for a value of Λ exceeding 1.

A related question we ask is whether the second best induces more agglomeration than the first best. This turns out to be true because

$$\Lambda - \Theta = \frac{(4b+d)\,[4(a-bc) - 2bt + 3\theta b]}{2d\theta(8b+3d)},$$

which is negative, for $4(a - bc) > 2bt$ by (8.28). This implies that the planner's best response to the loss of control on prices is to take advantage of the localization economies by having more firms in the large cluster. Hence, we have the following:

Proposition 8.9 *When there is partial agglomeration, the second best involves more asymmetry between clusters than the first best.*

This surprising result is to be understood as follows. The planner's objective is now to damper too high prices by controlling locations, especially when varieties are very differentiated. To achieve this goal, the planner chooses to expand the cluster in the larger region because, in so doing, price competition is intensified and localization economies are made stronger. This confirms once more that price competition is a strongly dispersion force.

The results obtained in this section confirm and extend those derived in the previous section. In particular, we have shown that asymmetric clusters are likely to emerge when transport costs become sufficiently low. Moreover, forcing a more equalitarian distribution of firms is likely to increase the inefficiency of the economy because price competition acts as a fairly strong dispersion force. Such results run against the conventional wisdom and point to a potential conflict between efficiency and spatial equity in the spatial distribution of production.

8.5 CONCLUDING REMARKS

The different models studied in this chapter allow us to shed some light on the main principles governing the spatial organization of a multiregional economy. In this perspective, one of the most striking results is that *a small initial advantage may lead to the emergence of a strongly polarized space* once we explicitly account for the existence of localized production externalities, natural amenities, or both. This effect is magnified when the mobility of factors or the transportability of products are high, or both. Indeed, either of the first two possibilities allows the localized externalities to display their full impact.

In addition, although they are active at the local level only, localization economies potentially appear to be a strong agglomeration force in shaping the interregional landscape (think of the Silicon Valley). This is especially true when transport costs between regions are low, varieties are very differentiated, or both. In these cases, agglomeration occurs because firms are able to enjoy a higher level of localization economies while they are still able to sell a substantial fraction of their output on distant markets. As in Chapters 6 and 7, in which we dealt with urban landscapes, low transport costs yield more agglomeration, but now this occurs at the interregional level. Our analysis has also identified a reason explaining how and when a firm may benefit from the presence of local competitors. Another reason will be discussed in the next chapter.

That more product differentiation fosters more agglomeration is not new. This fact agrees with what we have seen in Chapter 7 as well as with what we will see in the next chapter. The reason is always the same: a higher degree of product differentiation allows firms to relax price competition, thus permitting any existing agglomeration force to dominate the dispersion force. This force appears to be crucial in very different spatial settings and is likely to be strong in modern economies. Thus, the same causes lead to the same effects, although the geographical scale as well as the forces at work are very different. Interestingly, the same tendency holds at the first best optimum.

When the desirability for the output of the industry rises, more firms tend to locate within the same cluster, whose relative size increases at the expense of the other firms. Consequently, economic growth, as measured by the relative importance of the differentiated good to the numéraire, tends to foster more geographical concentration of production.[17] It is worth pointing out that such an increase in the agglomeration of firms arises even though the spatial distribution of demand remains unchanged. However, when more firms enter as a result of growth, the impact on agglomeration depends on the many features of the economy.

Unlike what many regional analysts and planners would argue, the optimal configuration tends to involve a more unbalanced distribution of firms than the market outcome, both in the duopoly and large group cases. If localization economies become increasingly important in advanced economic sectors, as suggested by the growing role played by knowledge spillovers in research and development, the observed regional imbalances in the geographical distribution of high-tech activities may not signify a wasteful allocation of resources. On the contrary, the size of the existing clusters could well be too small.

In this chapter, we have stressed the role of nonmarket interactions in the formation of industrial clusters. Without such technological externalities ($E(H_r) = 1$ in 8.2, $\theta_r = 0$ in 8.3, and $\theta = 0$ in 8.4), the models studied would always have generated the same outcome, namely, dispersion. Therefore, for the settings considered in this chapter, these externalities have been critical in showing the emergence of industrial agglomerations.

In the next chapter, we will move to the study of pecuniary externalities. Several reasons explain this alternative research strategy. First, even though technological and pecuniary externalities are natural components of any complete explanation of economic clustering, one major intellectual advantage is presented by the latter. Because they rely on market interactions, the origin of pecuniary externalities is clearer. In contrast, as discussed in Chapter 1, technological externalities often have the nature of a black box, thus leading to results that may have a vague meaning. In addition, although only technological externalities affect efficiency in an otherwise perfect market, with imperfect competition pecuniary externalities also matter in that changes in prices affect the deadweight loss due to existing distortions. Hence, both types of externalities matter when evaluating the efficiency of the market outcome. However, the inefficiencies associated with pecuniary externalities can be better understood because their origin is easier to identify. Last, it should be kept in mind that modeling Marshallian external effects by means of a technological externality may lead to inappropriate policy recommendations because the "proxy" is too rough (Abel-Rahman and Fujita 1990).

APPENDIX

A. We study the examples presented in Sections 8.2.1 and 8.2.2 in which we have $E(H_r) = \exp(\varepsilon H_r)$, $f(H_r) = H_r^\alpha$, $u(w_r^h) = \log w_r^h$ and $e(P_r) = aP_r - b(P_r/S)^2$ for $r = A, B$. Because the model of Section 8.2.1 is a special case of the one of 8.2.2 in which both a and b are set equal to zero, it is sufficient to consider the general case.

Using the expressions above, we can readily verify that

$$\dot{H}_A = \psi(H_A) = \log\left(w_A^H / w_B^H\right) + e(H_A + 1) - e(H - H_A + 1).$$

Because $w_r^H = \exp(\varepsilon H_r) f'(H_r) = \alpha \exp(\varepsilon H_r) H_r^{\alpha-1}$ for $r = A, B$, we have

$$\psi(H_A) = \varepsilon(2H_A - H) - (1-\alpha)\log[H_A/(H - H_A)] \\ + e(H_A + 1) - e(H - H_A + 1).$$

Differentiating this expression with respect to H_A yields

$$\psi'(H_A) = 2(\varepsilon + a) - (1-\alpha)/H_A - (1-\alpha)/(H - H_A) \\ - 2b(H+2)/S^2, \tag{A.1}$$

whose derivative is

$$\psi''(H_A) = (1-\alpha)\big/H_A^2 - (1-\alpha)/(H - H_A)^2.$$

Because $\psi'''(H_A)$ is strictly negative for $0 < H_A < H$, the function $\psi''(H_A)$ is strictly decreasing on $(0, H)$. Moreover, we have $\psi''(0) = \infty$ and $\psi''(H) = -\infty$. Therefore, there is a unique value $H_A^* = H/2$ such that $\psi''(H_A) = 0$. In other words, H_A^* is the unique maximizer of $\psi'(H_A)$ over $[0, H]$.

Two cases may then arise:

1. $\psi'(H_A^*) \leq 0$. This implies that we must have $\psi'(H_A) \leq 0$ for all values of H_A in $[0, H]$. This in turn implies that ψ is decreasing over the interval $(0, H)$. Because $\psi(0) = \infty$ and $\psi(H) = -\infty$, there is a unique equilibrium given by $H/2$, and this equilibrium is globally stable because $\psi'(H_A) < 0$ in a neighborhood of H_A^* (see Figure 8.1).
2. $\psi'(H_A^*) > 0$. Because $\psi'(H_A)$ is concave, two values H_A^- and H_A^+ exist for which $\psi'(H_A) = 0$. As a result, $\psi(H_A)$ is strictly decreasing over $[0, H_A^-)$, strictly increasing over (H_A^-, H_A^+), and strictly decreasing over $(H_A^+, H]$. Consequently, three equilibria exist if and only if $\psi(H_A^-) < \psi(H/2) < \psi(H_A^+)$ (see Figure 8.2).

Because regions have identical endowments, (A.1) becomes

$$\psi'(H_A^*) = \psi'(H/2) = 2(\varepsilon + a) - 4(1-\alpha)/H - 2b(H+2)/S^2,$$

and thus the condition for the existence of three equilibria $\psi'(H_A^*) > 0$ is equivalent to

$$\varepsilon + a - b(H+2)/S^2 > 2(1-\alpha)/H,$$

which is exactly (8.13). When this inequality is reversed, there is a single equilibrium because $\psi'(H_A^*) \leq 0$.

B. In the case of two varieties, we know that the quadratic utility is given by (8.14):

$$U(q_1, q_2) = \alpha(q_1 + q_2) - (\beta/2)\big(q_1^2 + q_2^2\big) - \delta q_1 q_2 + z.$$

In the case of $n > 2$ varieties, (8.14) may be extended as follows:

$$\begin{aligned}
U(q) &= \alpha \sum_{i=1}^{n} q_i - (\beta/2) \sum_{i=1}^{n} q_i^2 - (\delta/2) \sum_{i=1}^{n} \sum_{j \neq i}^{n} q_i q_j + z \\
&= \alpha \sum_{i=1}^{n} q_i - [(\beta - \delta)/2] \sum_{i=1}^{n} q_i^2 - (\delta/2) \sum_{i=1}^{n} \left(q_i \sum_{j=1}^{n} q_j \right) + z \\
&= \alpha \sum_{i=1}^{n} q_i - [(\beta - \delta)/2] \sum_{i=1}^{n} q_i^2 - (\delta/2) \sum_{i=1}^{n} \sum_{j=1}^{n} q_i q_j + z \\
&= \alpha \sum_{i=1}^{n} q_i - [(\beta - \delta)/2] \sum_{i=1}^{n} q_i^2 - (\delta/2) \left(\sum_{i=1}^{n} q_i \right)^2 + z.
\end{aligned}$$

Letting $n \to \infty$ and $q_i \to 0$, we then obtain (8.25) in which M stands for the mass of varieties.

NOTES

1. Not surprisingly, Marshallian externalities are the engine of economic development in the new theories of growth (Romer 1986; Lucas 1988).
2. Remember that Chapter 6 focused on the informational spillover interpretation only.
3. For recent contributions, see Glaeser et al. (1992); Henderson, Kuncoro, and Turner (1995); Ciccone and Hall (1996); Hanson (1996); Feldman and Audretsch (1999); and Smith (1999). Eberts and McMillen (1999) provide a regional science-oriented survey. In the same vein, Lucas (2001) has argued that local interactions are a basic component of the engine of growth.
4. According to Prescott (1998), each place is characterized by a "social capital" affecting the overall productivity; this capital will vary from place to place.
5. The content of that section is also related to the brain drain problem (see, e.g., Miyagiwa 1991), but we do not address this issue here.
6. Note also that Black (2000) has argued in a convincing way that, in the presence of mobility costs, skilled workers have incentives to move before unskilled workers do.
7. The neoclassical model based on wage differentials has been revisited by Rosen (1979) to account for the importance of regional amenities in the decision to migrate. Subsequently, his analysis was extended by Robak (1982), who added firms to the model to permit individuals to be both consumers and workers.

8. The role of these two types of externalities has been emphasized in the new growth theories and is at the heart of urban economics.
9. Assuming that both prices are identical and equal to p, we see that the aggregate demand for the differentiated product equals $2(a - bp)$, which is independent of d. Hence (8.15) has the desirable property that the market size in the industry does not change when the substitutability parameter d varies. More generally, it is possible to decrease (increase) d through a decrease (increase) in the parameter δ in the utility U keeping the other structural parameters a and b of the demand system unchanged.
10. Given (8.19), arbitrage is never profitable.
11. If $\theta_A = \theta_B$, the second case never arises, and there are only two equilibria.
12. The results presented in this section have been extended to the case of several firms and districts by Soubeyran and Weber (2002), considering a Cournot framework and an integrated market.
13. This agrees with the well-known idea that firms may be locked into an inferior technology (Arthur 1994, chap. 2).
14. A similar conflict appears in a different but related context in which the relocation of firms is triggered by a wage differential (Cordella and Grilo 2001).
15. See Part B of the chapter appendix for a justification of the coefficients used in that utility function.
16. The intuition behind this interpretation is very similar to the one that stands behind the Herfindahl index used to measure industrial concentration. If we control for the total amount of the differentiated good consumption, the absolute value of the quadratic term in (8.25) increases with the concentration of consumption on a few varieties, thus decreasing utility.
17. Compare (8.15) and (8.27).
18. Similarly, we have $\alpha = a/b$, $\beta = 1/(b + d)$, and $\delta = d/b(b + d)$. Hence, there is a one-to-one correspondence between the two sets of parameters.
19. This agrees with the analysis of the interplay between agglomeration and growth developed by Martin and Ottaviano (1999).

9

Industrial Agglomeration under Monopolistic Competition

9.1 INTRODUCTION

The spatial economy is replete with *pecuniary externalities*. For example, when some workers (firms) choose to migrate, they are likely to affect both the labor and product markets in their region of origin as well those in their region of destination. Such pecuniary externalities are especially relevant in the context of imperfectly competitive markets because prices do not perfectly reflect the social values of individual decisions. Pecuniary externalities are also better studied within a general equilibrium framework to account for the interactions between the product and labor markets. Among other things, this allows one to study the dual role of individuals as workers and consumers. At first sight, this seems to be a formidable task. Yet, as shown by Krugman (1991a), several of these various effects can be combined and studied within a relatively simple general equilibrium model of monopolistic competition, which has come to be known as the *core–periphery model*.[1]

Recall that monopolistic competition à la Chamberlin (1933) involves consumers with a preference for variety (*varietas delectat*), whereas firms producing these varieties compete for a limited amount of resources because they face increasing returns. The prototype that has emerged from the industrial organization literature is the model developed by Spence (1976) and Dixit and Stiglitz (1977), which is sometimes called the S–D–S model. These authors assumed that each firm is negligible in the sense that it may ignore its impact on, and hence reactions from, other firms, but retains enough market power for pricing above marginal cost regardless of the total number of firms (like a monopolist). Moreover, the position of a firm's demand depends on the actions taken by all firms in the market (as in perfect competition).

In many applications, the S–D–S model has proven to be a very powerful instrument for studying the aggregate implications of monopoly power and increasing returns – especially so when these elements are the basic ingredients of self-sustaining processes such as those encountered in modern theories of growth and economic geography (Matsuyama 1995). This is so because of the following reasons:

1. Although each firm is a price-maker, strategic interactions are very weak in this model, thus making the existence of an equilibrium much less problematic than in general equilibrium under imperfect competition (see, e.g., Bonanno 1990).
2. The assumption of free entry and exit leads to zero profit so that a worker's income is just equal to his wage, which is another major simplification.[2]
3. Although it is seldom noticed, the difference between price competition and quantity competition becomes immaterial in monopolistic competitive settings. Indeed, being negligible to the market, each firm behaves as a monopolist on the residual demand, which makes it indifferent between using price or quantity as a strategy.[3]

The setting considered by Krugman, described in Section 9.2, is simple but rich enough to capture the major forces at work in the multiregional economy. It combines the Dixit–Stiglitz model with the iceberg-type transport cost in which, as in Thünen, only a fraction of the good shipped reaches its destination (Samuelson 1954a; 1983). There are two sectors: (1) the agricultural sector producing a homogeneous good under constant returns to scale and (2) the manufacturing sector supplying a differentiated good under increasing returns. As usual, the market equilibrium is the outcome of the interplay between a dispersion force and an agglomeration force. The centrifugal force is simple. It lies in the following two sources: (1) the spatial immobility of farmers whose demands for the manufactured good are to be met and (2) the increasing competition that arises when firms are more agglomerated. The centripetal force is more involved. If a larger number of manufactures are located in one region, the number of varieties locally produced is also larger. Then, because firms do not price discriminate between regions, the equilibrium price index of manufactured goods is lower in this region. This in turn, induces some workers living in the smaller region, to move toward the larger region, where they may enjoy a higher standard of living. The resulting increase in the numbers of workers creates a larger demand for the differentiated good, which, therefore, leads additional firms to locate in this region. This implies the availability of more varieties in the region in question but less in the other because there are scale economies at the firm's level. Consequently, as noticed by Krugman (1991a, 486), *circular causation* à la Myrdal is present because these two effects reinforce each other: "manufactures production will tend to concentrate where there is a large market, but the market will be large where manufactures production is concentrated."

When transportation costs (or, more generally, trade costs) are sufficiently low, Krugman (1991a) has shown that all manufactures are concentrated in a single region that becomes the *core* of the economy, whereas the other region, called the *periphery*, supplies only the agricultural good. Firms are able to exploit increasing returns by selling more in the larger market without losing much business in the smaller market. For exactly the opposite reason, the economy displays a symmetric regional pattern of production when transportation costs are high. Hence, this model allows for the possibility of divergence between regions, whereas the neoclassical model, based on constant returns and perfect competition in the two sectors, would predict convergence only. As in previous chapters, there would be a monotone decreasing relationship between the degree of agglomeration and the level of transport costs. In addition, Krugman has also shown that there are often multiple equilibria; thus, there is a need for some refinement. As before, instability is used to dismiss some equilibria. Several stable equilibria may exist; consequently, the market outcome is likely to depend on the initial conditions. Or, to put it differently, history matters.[4] The process of circular causation, then, looks like a snowball effect that leads industrial firms to be locked in the same region for long periods: think of the Industrial Belt in the United States, the Pacific Industrial Belt in Japan, or the "Blue Banana" in Europe. This result also represents a neat formalization of an idea put forward by Myrdal (1957, 26–7) a long time ago:

> Within broad limits, the power of attraction of a centre has its origin mainly in the historical accident that something was once started there, and not in a number of other places where it could equally well have been started, and that start met success.

In such a context, it is well known that small initial differences between territories, minor changes in the socioeconomic environment, or both may eventually result in vastly different economic configurations.[5]

Note that the core–periphery model can be used to study the growing geographical concentration of business services in large metropolitan areas (Kolko 1999). In this context, the manufacturing sector is to be reinterpreted as the *service industry* in which firms not only supply consumers and manufacturing firms but also serve each other. This circularity in demand gives the corresponding firms incentives to agglomerate that are very similar to those studied by Krugman. As a result, once the different sectors are properly reinterpreted to account for the new trends emerging in developed economies, the core–periphery model will keep its relevance despite the gradual shift of manufacturers toward smaller cities or even rural areas. This tendency will be strengthened by the growing inclination of business services to work near the headquarters and research laboratories of manufacturing firms, which remain mostly located in large cities.

In Section 9.3, we formalize this idea by considering a large array of more or less tradable intermediate goods (and business services) as a major cause for agglomeration. So far, indeed, agglomeration has been considered as the outcome of a circular causation process fed by the mobility of workers. Yet, in the international marketplace, it is reasonable to expect this mobility to be low, and thus the core–periphery model cannot be used to explain the agglomeration of industries in the world economy. Consequently, as argued by Venables (1996), dealing with the intermediate sector allows us to explain the possible emergence of a core–periphery structure at the international level. However, instead of following Venables (who assumes that both the intermediate and final sectors operate under increasing returns and monopolistic competition), we use a simpler framework. As in Section 4.2.1, we assume that the intermediate sector produces a differentiated good and exhibits increasing returns; however, the final sector produces a homogeneous good and exhibits constant returns. Finally, it is supposed that workers remain in their region. We then show that both sectors concentrate within the same region provided that the transport costs of the intermediate goods are sufficiently high (typically when they are nontradable). This is so even when the transport cost of the final good is very low. Indeed, the agglomeration of the intermediate sector firms makes it profitable for the final sector firms to agglomerate with them despite the wage gap generated by the immobility of workers.

In summary, the core–periphery structure may emerge owing to the migration of workers and the imperfectly competitive nature of the final sector or to the existence of an imperfectly competitive intermediate sector when workers are immobile. This result is very important for the space-economy, and thus it is crucial to know how it depends on the specificities of the framework employed. First, the use of the CES utility and iceberg cost leads to a convenient setting in which demands have a constant elasticity. However, such a result conflicts with research in spatial pricing theory in which demand elasticity varies with distance. Second, although the iceberg cost is able to capture the fact that shipping is resource-consuming, such a modeling strategy implies that any increase in the mill price is accompanied with a proportional increase in transport cost, which often seems unrealistic.[6] Finally, although models are based on very specific assumptions, they are often beyond the reach of analytical resolution, forcing authors to appeal to numerical investigations.[7] As recognized by Krugman (1998, 164) himself, "To date, the new economic geography has depended heavily on the tricks summarized in Fujita, Krugman, and Venables (1999) with the slogan 'Dixit–Stiglitz, iceberg, evolution and the computer.' "

This state of affairs has led Ottaviano and Thisse (1998) and Ottaviano, Tabuchi, and Thisse (2002) to revisit the core–periphery model using an alternative framework that involves downward-sloping linear demands and a linear transport cost measured in terms of the numéraire. Such a setting, which is very popular in location theory (Beckmann and Thisse 1986; Greenhut, Norman, and Hung 1987), takes us far away from the model used by Krugman and offers the advantage of yielding analytical solutions. Although the conclusions are not exactly the same as those derived by Krugman, this alternative model also yields a core–periphery structure once transportation costs are sufficiently low. Therefore, the core–periphery structure seems to be robust against very different formulations of preferences and transport technologies.

The linear model permits the study of different spatial price policies (Ottaviano and Thisse 1998). Because mill pricing yields the same qualitative results as in Section 9.2, we restrict ourselves to the case of discriminatory pricing in Section 9.4. This framework is very simple to use and is also suitable for studying the welfare properties of the core–periphery structure – an issue that has been untouched in most economic geography models. In Section 9.5, using the model of Section 9.4, we focus on the interplay between history and expectations in the formation of the economic space when migrants maximize the intertemporal value of their utility flows.

The work of Krugman has triggered a plethora of contributions, which have been surveyed by Ottaviano and Puga (1998). As noted earlier, the main result obtained by Krugman is the monotone relationship between the degree of agglomeration and the transportation cost level. In Section 9.6, the generality of such a relationship is discussed through several modifications of the basic model.

9.2 THE CORE–PERIPHERY MODEL

Although we focus on a two-region economy in this chapter, it will prove convenient to have a more general framework for subsequent developments.

9.2.1 The Framework

The economic space is made of R regions. The economy has two sectors: the modern sector ($\mathbb{M}$) and the traditional sector ($\mathbb{T}$). There are two production factors: the high-skilled workers and the low-skilled workers. The $\mathbb{M}$-sector produces a continuum of varieties of a horizontally differentiated product under increasing returns using skilled labor as the only input. The $\mathbb{T}$-sector produces a homogeneous good under constant returns using unskilled labor as the only input.

The economy is endowed with L unskilled workers and with H skilled workers. The skilled workers are perfectly mobile between regions, whereas the unskilled are immobile. As discussed in Section 8.2, this extreme assumption is partially justified because the skilled are more mobile than the unskilled. The share of unskilled workers in region r is fixed and denoted $0 \leq \nu_r \leq 1$ for $r = 1, \ldots, R$. The share of skilled workers in each region r is variable and denoted by $0 \leq \lambda_r \leq 1$ for $r = 1, \ldots, R$.

Although both consumption and production take place in a specific region, it is notationally convenient to describe preferences and technologies without explicitly referring to any particular region.

Preferences are identical across all workers and described by a Cobb–Douglas utility,

$$U = Q^{\mu} T^{1-\mu} / \mu^{\mu} (1-\mu)^{1-\mu} \qquad 0 < \mu < 1, \tag{9.1}$$

where Q stands for an index of the consumption of the modern sector varieties, and T is the consumption of the output of the traditional sector. When the modern sector provides a continuum of varieties of size M, the index Q is given by

$$Q = \left[\int_0^M q(i)^\rho di \right]^{1/\rho} \qquad 0 < \rho < 1, \tag{9.2}$$

where $q(i)$ represents the consumption of variety $i \in [0, M]$. Hence, each consumer displays a preference for variety. In (9.2), the parameter ρ stands for the inverse of the intensity of desire for variety over the differentiated product. When ρ is close to 1, varieties are close to perfect substitutes; when ρ decreases, the desire to spread consumption over all varieties increases. If we set

$$\sigma \equiv \frac{1}{1-\rho},$$

then σ is the elasticity of substitution between any two varieties, which varies between 1 and ∞. Because there is a continuum of firms, each firm is negligible, and the interactions between any two firms are zero, but aggregate market conditions (e.g., the average price across firms) affect each firm. This provides a setting in which firms are not competitive (in the classic economic sense of having infinite demand elasticity), but at the same time they have no strategic interactions with one another (see (9.4) below).[8]

If Y denotes the consumer income, $p^{\mathbb{T}}$ the price of the traditional good, and $p(i)$ the price of variety i, then the demand functions are

$$T = (1-\mu)Y/p^{\mathbb{T}} \tag{9.3}$$

$$q(i) = \frac{\mu Y}{p(i)} \frac{p(i)^{-(\sigma-1)}}{P^{-(\sigma-1)}} = \mu Y p(i)^{-\sigma} P^{\sigma-1} \qquad i \in [0, M], \tag{9.4}$$

where P is the price index of the differentiated product given by

$$P \equiv \left[\int_0^M p(i)^{-(\sigma-1)} di \right]^{-1/(\sigma-1)}. \tag{9.5}$$

Introducing (9.3) and (9.4) into (9.1) yields the indirect utility function

$$v = Y P^{-\mu} (p^{\mathbb{T}})^{-(1-\mu)}. \tag{9.6}$$

The technology in the $\mathbb{T}$-sector is such that one unit of output requires one unit of L. Each variety of the $\mathbb{M}$-sector is produced according to the same technology such that the production of the quantity $q(i)$ requires $l(i)$ units of skilled labor given by

$$l(i) = f + cq(i), \tag{9.7}$$

where f and c are, respectively, the fixed and marginal labor requirements. Clearly, this technology exhibits scale economies. Without loss of generality, we choose the unit of skilled labor such that $c = 1$. Because preferences exhibit love for diversity and there are increasing returns but no scope economies, each variety is produced by a single firm. Indeed, any firm obtains a higher share of the market by producing a differentiated variety than by replicating an existing one.[9] In turn, this implies that the mass of firms is identical to the mass of varieties and that the output of a firm equals the demand of the corresponding variety.

The output of the $\mathbb{T}$-sector is costlessly traded between any two regions and is chosen as the numéraire, and thus $p^{\mathbb{T}} = 1$. In contrast, the output of the $\mathbb{M}$-sector is shipped at a positive cost according to the "iceberg" technology: when one unit of the differentiated product is moved from region r to region s, only a fraction $1/\Upsilon_{rs}$ arrives at destination, where $\Upsilon_{rs} > 1$ for $r \neq s$ and $\Upsilon_{rr} = 1$. Hence, if variety i is produced in region r and sold at the mill (fob) price $p_r(i)$, the price $p_{rs}(i)$ paid by a consumer located in region s ($\neq r$) is

$$p_{rs}(i) = p_r(i)\Upsilon_{rs}. \tag{9.8}$$

If the distribution of firms is $(M_1, \ldots, M_R)$, using (9.5) and setting $\Upsilon_{rr} = 1$, we obtain the price index P_r in region r from

$$P_r = \left\{ \sum_{s=1}^{R} \Upsilon_{sr}^{-(\sigma-1)} \int_0^{M_s} p_s(i)^{-(\sigma-1)} di \right\}^{-1/(\sigma-1)}. \tag{9.9}$$

Let w_r denote the wage rate of a skilled worker living in region r. Because the price of the traditional good equals 1, the wage of the unskilled workers is also equal to 1 in all regions. Thus, because there is free entry and exit, and therefore zero profit in equilibrium, the income of region r is

$$Y_r = \lambda_r H w_r + \nu_r L. \tag{9.10}$$

From (9.4), the total demand of the firm producing variety i and located in region r is

$$\begin{aligned} q_r(i) &= \sum_{s=1}^{R} \mu Y_s [p_r(i)\Upsilon_{rs}]^{-\sigma} (P_s)^{\sigma-1} \Upsilon_{rs} \\ &= \mu p_r(i)^{-\sigma} \sum_{s=1}^{R} Y_s \Upsilon_{rs}^{-(\sigma-1)} (P_s)^{\sigma-1}. \end{aligned} \tag{9.11}$$

This expression requires some comment. The term $\mu Y_s [p_r(i)\Upsilon_{rs}]^{-\sigma} (P_s)^{\sigma-1} \Upsilon_{rs}$ stands for the quantity shipped from the firm located in r to region s. Here, the regional consumption in s, which is equal to $\mu Y_s [p_r(i)\Upsilon_{rs}]^{-\sigma} (P_s)^{\sigma-1}$, must be multiplied by Υ_{rs} because the firm's output "melts" on the way, thus implying that the firm must send out a larger quantity of its output for the desired quantity to be delivered.[10]

Because each firm has a negligible impact on the market, it may accurately neglect the impact of a price change over consumers' income (Y_r) and other firms' prices and hence on the regional price indices (P_r).[11] Consequently, (9.11) implies that, regardless of the spatial distribution of consumers, each firm faces an isoelastic downward-sloping demand (the elasticity equals σ). This very convenient property depends crucially on the assumption of an iceberg transport cost, which affects the level of demand but not its elasticity.

The profit function of a firm in r is

$$\pi_r(i) = p_r(i)q_r(i) - w_r[f + q_r(i)] = [p_r(i) - w_r]q_r(i) - w_r f. \quad (9.12)$$

Because varieties are equally weighted in the utility function, the equilibrium price is the same across all firms located in region r. Solving the first-order condition using (9.11) yields the common equilibrium price

$$p_r^* = \frac{w_r}{\rho} \quad r = 1, \ldots, R. \quad (9.13)$$

This means that firms uses a relative markup equal to $1/\rho$, which is independent of the firms' and consumers' distributions. Everything else being equal, more product differentiation leads to a higher markup and, therefore, to a higher equilibrium price. However, the equilibrium price depends on the mass of firms and workers established in region r through the local wage w_r.

Substituting (9.13) into the profit function leads to

$$\pi_r = \frac{w_r}{\sigma - 1}q_r - w_r f = \frac{w_r}{\sigma - 1}[q_r - (\sigma - 1)f]. \quad (9.14)$$

Under free entry, profits are zero, and thus the equilibrium output of a firm is a constant given by

$$q_r^* = (\sigma - 1)f \qquad r = 1, \ldots, R. \quad (9.15)$$

Note that this quantity is independent of the distributions of firms and workers and is the same across regions. As a result, in equilibrium a firm's labor requirement is also unrelated to the firms' distribution:

$$l^* = \sigma f \qquad r = 1, \ldots, R.$$

Thus, the total mass of firms in the $\mathbb{M}$-sector is constant and equal to H/l^*, whereas the corresponding firm distribution

$$M_r = \lambda_r H/l^* = \lambda_r H/\sigma f \qquad r = 1, \ldots, R \quad (9.16)$$

depends only on the distribution of skilled workers. These equalities imply that the core–periphery model allows for the spatial redistribution of the modern sector but not for its growth, for the total number of firms (or varieties) is constant; this issue is addressed in Chapter 11.

Introducing the equilibrium prices (9.13) and substituting (9.16) for M_r in the regional price index (9.9), we obtain

$$P_r = \left[\sum_{s=1}^{R} \frac{\lambda_s H}{\sigma f} \left(\frac{w_s}{\rho} \Upsilon_{sr} \right)^{-(\sigma-1)} \right]^{-1/(\sigma-1)}$$

$$= \kappa_1 \left[\sum_{s=1}^{R} \lambda_s (w_s \Upsilon_{sr})^{-(\sigma-1)} \right]^{-1/(\sigma-1)} \qquad r = 1, \ldots, R, \qquad (9.17)$$

where

$$\kappa_1 \equiv \rho^{-1} \left(\frac{H}{\sigma f} \right)^{-1/(\sigma-1)},$$

which clearly depends on the spatial distribution of skilled workers as well as on the values of transport costs.

Finally, we consider the labor market clearing conditions for a given distribution of workers. The wage prevailing in region r is the highest wage that firms located there can pay under the nonnegative profit constraint. For that, we evaluate the demand (9.11) as a function of the wage through the equilibrium price (9.13):

$$q_r(w_r) = \mu \left(\frac{1}{\rho} \right)^{-\sigma} w_r^{-\sigma} \sum_{s=1}^{R} Y_s \Upsilon_{rs}^{-(\sigma-1)} P_s^{\sigma-1}. \qquad (9.18)$$

Because this expression is equal to $(\sigma - 1) f$ when profits are zero, we obtain the following implicit expression for the *zero-profit wages*:

$$w_r^* = \kappa_2 \left[\sum_{s=1}^{R} Y_s \Upsilon_{rs}^{-(\sigma-1)} P_s^{\sigma-1} \right]^{1/\sigma} \qquad r = 1, \ldots, R, \qquad (9.19)$$

where

$$\kappa_2 \equiv \rho[\mu/(\sigma - 1) f]^{1/\sigma}.$$

Clearly, w_r^* is the equilibrium wage prevailing in region r when $\lambda_r > 0$.

Substituting (9.19) for Y and setting $p^{\mathbb{T}} = 1$ in the indirect utility (9.6), we obtain the *real wage* in region r as follows:

$$v_r = \omega_r = \frac{w_r^*}{P_r^{\mu}} \quad r = 1, \ldots, R. \qquad (9.20)$$

Hence, the indirect utility is here equivalent to maximizing the real wage.

Finally, the Walras law implies that the traditional sector market is also in equilibrium provided that the equilibrium conditions above are satisfied.

For a given spatial distribution of skilled workers, we now ask whether there is an incentive for them to migrate and, if so, what direction the flow of migrants will take. A *spatial equilibrium* arises when no skilled worker may get a higher utility level in another region: $(\lambda_1^*, \ldots, \lambda_R^*)$ is a spatial equilibrium if there exists a positive constant ω^* such that

$$\omega_r \leq \omega^* \quad \text{for } r = 1, \ldots, R$$
$$\omega_r = \omega^* \quad \text{if } \lambda_r^* > 0.$$

Hence, the zero-profit real wage that local firms could afford to pay in a region containing no skilled workers is lower than (or just equal to) the equilibrium real wage. Because the functions $\omega_r(\lambda_1, \ldots, \lambda_R)$ are continuous in $(\lambda_1, \ldots, \lambda_R)$ over the compact set

$$\Lambda \equiv \left\{ (\lambda_1, \ldots, \lambda_R); \sum_{r=1}^{R} \lambda_r = 1 \quad \text{and} \quad \lambda_r \geq 0 \right\},$$

we can appeal to Proposition 1 of Ginsburgh, Papageorgiou, and Thisse (1985) to guarantee that such an equilibrium always exists.

Following a now well-established tradition in migration modeling, we focus on an adjustment process in which workers are attracted (repulsed) by regions providing high (low) utility levels:

$$\dot{\lambda}_r = \lambda_r(\omega_r - \bar{\omega}) \qquad r = 1, \ldots, R,$$

where $\dot{\lambda}_r$ is the time-derivative of λ_r, ω_r is the equilibrium real wage corresponding to the distribution $(\lambda_1, \ldots, \lambda_R)$, and $\bar{\omega} \equiv \sum \lambda_s \omega_s$ is the average real wage across all regions. In other words, the skilled move from the low-wage regions toward the high-wage ones.

A spatial equilibrium is stable if, for any marginal deviation of the population distribution from the equilibrium, the equation of motion above brings the distribution of skilled workers back to the original one. In doing so, we assume that local labor markets adjust instantaneously when some skilled workers move from one region to another. More precisely, the mass of firms in each region must be such that the labor market clearing conditions (9.16) remain valid for the new distribution of workers. Wages are then adjusted in each region for each firm to earn zero profits in any region having skilled workers because workers move toward high-wage regions.

Observe here one more justification for working with a continuum of agents (workers and firms): this modeling strategy allows one to respect the integer nature of a worker's or firm's location while describing the evolution of the regional share of production by means of differential equations.

9.2.2 The Two-Region Case

Consider two regions A and B. The unskilled workers are equally split between regions ($\nu_A = \nu_B = 1/2$). To keep things as symmetric as possible, we also assume that $\Upsilon_{AB} = \Upsilon_{BA} \equiv \Upsilon$. In this specific context, the basic equations developed in the foregoing are as follows:

$$Y_r = \lambda_r H w_r + L/2 \qquad r = A, B \tag{9.21}$$

$$P_r = \kappa_1 \left[\lambda_r w_r^{-(\sigma-1)} + \lambda_s (w_s \Upsilon)^{-(\sigma-1)}\right]^{-1/(\sigma-1)} \qquad s \neq r \tag{9.22}$$

$$w_r^* = \kappa_2 \left(Y_r P_r^{\sigma-1} + Y_s \Upsilon^{-(\sigma-1)} P_s^{\sigma-1}\right)^{1/\sigma} \qquad s \neq r \tag{9.23}$$

$$\omega_r = w_r^* P_r^{-\mu} \qquad r = A, B. \tag{9.24}$$

Whenever this turns out to be convenient, from now we use $\lambda \equiv \lambda_A$ so that $\lambda_B = 1 - \lambda$. Given a parametric solution to the system (9.21)–(9.24), a *spatial equilibrium* arises at $\lambda \in (0, 1)$ when

$$\Delta\omega(\lambda) \equiv \omega_A(\lambda) - \omega_B(\lambda) = 0$$

or at $\lambda = 0$ when $\Delta\omega(0) \leq 0$, or at $\lambda = 1$ when $\Delta\omega(1) \geq 0$.

The stability is studied with respect to the following equation of motion:

$$\dot{\lambda} = \lambda \Delta\omega(\lambda)(1 - \lambda) \tag{9.25}$$

and is defined as in Section 8.4.[12] If $\Delta\omega(\lambda)$ is positive and $\lambda \in (0, 1)$, workers move from B to A; if it is negative, they go in the opposite direction. Clearly, any spatial equilibrium is a steady-state for (9.25).

The system (9.21)–(9.24) of nonlinear equations cannot be solved analytically. As a consequence, deriving a characterization of its solution in terms of λ is not simple. To derive some insight into the nature of the equilibrium, computational experiments have been performed by Krugman (1991a).[13] The results are displayed in Figure 9.1, where the following results appear. For a large value of Υ ($= \Upsilon_1$), there is only one equilibrium corresponding to the full dispersion of the modern sector ($\lambda = 1/2$), which is stable. When Υ takes some intermediate value Υ_2, four more equilibria emerge that are all asymmetric. However, the two interior equilibria are unstable. Hence, three stable equilibria now exist: the symmetric configuration and the core–periphery structure with concentration of the modern sector in region A or region B. Finally, when Υ takes a sufficiently low value ($= \Upsilon_3$), the symmetric equilibrium becomes unstable, and thus the core–periphery structure is the only stable outcome.

These observations will serve as a guide in the rest of the analysis.

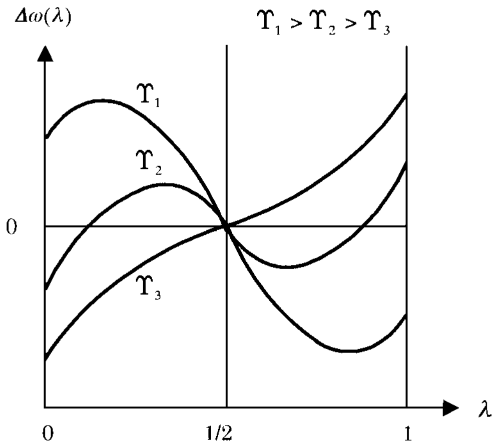


Figure 9.1: Migration dynamics under various values of Υ.

9.2.3 The Core–Periphery Structure

Suppose that the modern sector is concentrated in one region, say region A so that $\lambda = 1$. To check whether this is an equilibrium, we ask whether a skilled worker could be strictly better off in B. More precisely, we wish to determine conditions under which the real wage he may obtain in region B does not exceed the real wage this worker gets in region A. Setting $\lambda = 1$ in (9.21)–(9.24), we get the following equations:

$$Y_A = Hw_A^* + L/2 \quad \text{and} \quad Y_B = L/2$$
$$P_A = \kappa_1 w_A^* \quad \text{and} \quad P_B = \kappa_1 \Upsilon w_A^*. \tag{9.26}$$

Then, w_A^* is obtained by substituting (9.16) into (9.23) with $r = A$,

$$w_A^* = \kappa_2 \left[Y_A (\kappa_1 w_A^*)^{\sigma-1} + Y_B \Upsilon^{-(\sigma-1)} (\kappa_1 \Upsilon w_A^*)^{\sigma-1} \right]^{1/\sigma}$$

which yields $w_A^* = (\mu/H)(Y_A + Y_B)$, or

$$w_A^* = \frac{\mu}{1-\mu} \frac{L}{H}.$$

From (9.13), it is then possible to determine the common equilibrium price of all varieties in terms of the fundamentals of the economy:

$$p_A^* = \frac{1}{\rho} \frac{\mu}{1-\mu} \frac{L}{H},$$

which shows that the price of the differentiated product within the agglomeration increases with the unskilled–skilled ratio (L/H) as well as with the share of the modern sector (μ).

Finally, when the modern sector is geographically concentrated in A, the regional nominal incomes are as follows:

$$Y_A = \frac{\mu}{1-\mu}L + \frac{L}{2} \quad \text{and} \quad Y_B = \frac{L}{2},$$

and thus the gross domestic product (GDP) of the economy is given by

$$Y_{\mathrm{G}} \equiv Y_A + Y_B = L/(1-\mu).$$

The equilibrium real wage in region A is

$$\omega_A = \kappa_1^{-\mu}(w_A^*)^{1-\mu} = \left(\frac{1}{\rho}\right)^{-\mu}\left(\frac{H}{\sigma f}\right)^{\mu/(\sigma-1)}\left(\frac{\mu}{1-\mu}\frac{L}{H}\right)^{1-\mu},$$

which is independent of Υ.

Agglomeration in region A is an equilibrium if and only if ω_A is larger than or equal to ω_B. Thus, we need to determine ω_B. To find it, we substitute (9.22) for the price index and (9.23) for the nominal wage into the real wage (9.24) and get

$$\begin{aligned}\omega_B &= \kappa_1^{\rho-\mu}\kappa_2(w_A^*)^{\rho-\mu}\Upsilon^{-\mu}\big(Y_A\Upsilon^{-(\sigma-1)} + Y_B\Upsilon^{\sigma-1}\big)^{1/\sigma} \\ &= \kappa_1^{\rho-\mu}\kappa_2(w_A^*)^{\rho-\mu}\Upsilon^{-\mu}\big(Y_A\Upsilon^{-(\sigma-1)} + Y_B\Upsilon^{\sigma-1}\big)^{1/\sigma}.\end{aligned}$$

It can then readily be verified that

$$\frac{\omega_B}{\omega_A} = \left[\frac{1+\mu}{2}\Upsilon^{-\sigma(\mu+\rho)} + \frac{1-\mu}{2}\Upsilon^{-\sigma(\mu-\rho)}\right]^{1/\sigma}. \tag{9.27}$$

When shipping is costless ($\Upsilon = 1$), we always have $\omega_B/\omega_A = 1$: location does not matter. Furthermore, the first term in the right-hand side of (9.27) is always decreasing in Υ. Therefore, because the second term is also decreasing when $\mu \geq \rho$, the ratio ω_B/ω_A always decreases with Υ, thus implying that $\omega_B < \omega_A$ for all $\Upsilon > 1$. This means that the core–periphery structure is a stable equilibrium for all $\Upsilon > 1$. When

$$\mu \geq \rho,$$

which is called the *black hole condition*, varieties are so differentiated that firms' demands are not very sensitive to differences in transportation costs, thus making the agglomeration force very strong. In fact, it is so strong that agglomeration can be viewed as a "black hole" attracting any movable activity.

More interesting is the case in which

$$\mu < \rho, \tag{9.28}$$

that is, varieties are not very differentiated so that the firms' demands are sufficiently elastic and, hence, the agglomeration force is weaker. If (9.28)

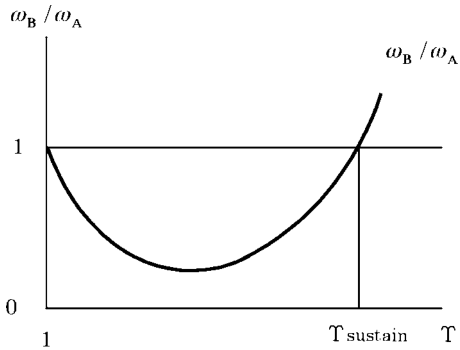


Figure 9.2: The determination of the sustain point.

holds, then the second term in (9.27) goes to infinity when $\Upsilon \to \infty$, and the ratio ω_B/ω_A is as depicted in Figure 9.2.[14]

We see that a single value $\Upsilon_{\text{sustain}} > 1$ exists such that $\omega_B/\omega_A = 1$. Hence, the agglomeration is a stable equilibrium for any $\Upsilon \leq \Upsilon_{\text{sustain}}$. In other words, once all firms belonging to the modern sector locate together within a region, they stay there as long as carrying their output to the other region is sufficiently cheap.[15] This occurs because firms can enjoy all the benefits of agglomeration without losing much of their business in the other region. Such a point is called the *sustain point* because, once firms are fully agglomerated, they stay so for all smaller values of Υ.[16] On the other hand, when transportation costs are sufficiently high ($\Upsilon > \Upsilon_{\text{sustain}}$), firms lose much on their exports so that the core–periphery structure is no longer an equilibrium.

Summarizing these results, we have the following:

Proposition 9.1 *Consider a two-region economy.*

1. *If $\mu \geq \rho$, then the core–periphery structure is always a stable equilibrium.*
2. *If $\mu < \rho$, then a unique solution $\Upsilon_{\text{sustain}} > 1$ exists to the equation*

$$\frac{1+\mu}{2}\Upsilon^{-\sigma(\mu+\rho)} + \frac{1-\mu}{2}\Upsilon^{-\sigma(\mu-\rho)} = 1 \tag{9.29}$$

such that the core–periphery structure is a stable equilibrium for any $\Upsilon \leq \Upsilon_{\text{sustain}}$.

It is remarkable that $\Upsilon_{\text{sustain}}$ depends only on the degree of product differentiation (σ) and the share of the modern sector in consumption (μ).

Interestingly, Proposition 9.1 provides formal support of the claim made by Kaldor (1970, 241) more than 30 years ago:

When trade is opened up between them, the region with the more developed industry will be able to supply the need of the agricultural area of the other region on more favourable terms: with the result that the industrial centre of the second region will lose its market and will tend to be eliminated.

The proposition also supports the claim of Giersch (1949, 94), who observed more than half a century ago that

production would tend to be centered in those industrial countries which already provide large domestic markets before the formation of the federal state.

It is worth stressing that the agglomeration is obtained as the aggregate outcome of a handful of individual decisions: the skilled workers do not choose a priori to be (or not to be) together. They are brought together through individual decisions based on current market prices and wages.

9.2.4 The Symmetric Structure

What we have just seen suggests that the modern sector is geographically dispersed when transportation costs are high and when (9.28) holds. To check this conjecture, we consider the symmetric configuration ($\lambda = 1/2$). In this case, there are only four equilibrium conditions,

$$Y_A = Y_B = Y = (H/2)w^* + L/2,$$

where w^* is the common zero-profit wage prevailing at the symmetric configuration given by

$$\begin{aligned} w^* &= \kappa_2\left(Y P^{\sigma-1} + Y\Upsilon^{-(\sigma-1)} P^{\sigma-1}\right)^{1/\sigma} \\ &= \kappa_2(Y P^{\sigma-1})^{1/\sigma}\left(1 + \Upsilon^{-(\sigma-1)}\right)^{1/\sigma}, \end{aligned}$$

and the common price index is equal to

$$\begin{aligned} P &= \kappa_1\left[\frac{1}{2}(w^*)^{-(\sigma-1)} + \frac{1}{2}(w^*\Upsilon)^{-(\sigma-1)}\right]^{-1/(\sigma-1)} \\ &= \kappa_1 2^{1/(\sigma-1)} w^*\left(1 + \Upsilon^{-(\sigma-1)}\right)^{-1/(\sigma-1)}. \end{aligned}$$

The common real wage is

$$\omega = w^* P^{-\mu}.$$

Because $\omega_A = \omega_B = \omega$, the symmetric structure is a spatial equilibrium for all $\Upsilon > 1$.

For a given $\Upsilon > 1$, the symmetric equilibrium is stable (unstable) if the slope of $\Delta\omega(\lambda)$ is negative (positive) at $\lambda = 1/2$. Checking this condition requires fairly long calculations using all the equilibrium conditions. However, Fujita, Krugman, and Venables (1999, chap. 5) have shown the following results. First, when (9.28) does not hold, the symmetric equilibrium is always unstable. However, when (9.28) holds, this equilibrium is stable (unstable) if Υ

is larger (smaller) than some threshold value Υ_{break} given by

$$\Upsilon_{\text{break}} = \left[\frac{(\rho+\mu)(1+\mu)}{(\rho-\mu)(1-\mu)}\right]^{1/(\sigma-1)}, \tag{9.30}$$

which is clearly larger than 1. This is called the *break point* because symmetry between the two regions is no longer a stable equilibrium for lower values of Υ. It is interesting to note that Υ_{break} depends on the same parameters as $\Upsilon_{\text{sustain}}$. It is apparent from (9.30) that Υ_{break} is increasing with the share of the modern sector (μ) and with the degree of product differentiation ($1/\rho$).

Figure 9.3 represents all stable (unstable) equilibria by solid (broken) lines. It is shown in the appendix of this chapter that $\Upsilon_{\text{break}} < \Upsilon_{\text{sustain}}$. Hence, a domain of transport cost values exists over which there is multiplicity of equilibria, namely agglomeration and dispersion. More precisely, for $\Upsilon > \Upsilon_{\text{sustain}}$, the economy necessarily involves full dispersion. For $\Upsilon < \Upsilon_{\text{break}}$, the core–periphery structure always arises, the winning region depending on the initial conditions: the region with the initially larger share of the modern sector ends up with the whole share. Finally, for $\Upsilon_{\text{break}} \leq \Upsilon \leq \Upsilon_{\text{sustain}}$, both full agglomeration and full dispersion are stable equilibria. In the corresponding domain, the economy displays some hysteresis because full dispersion still prevails when transport costs fall below the sustain point while staying above the break point. Eventually, as transportation costs decrease sufficiently, sudden agglomeration of the modern sector arises.

9.2.5 A Model of the Modern Sector with Two Factors

It is reasonable to view the factor standing behind the fixed costs as skilled labor, especially when these costs correspond to research and development as well as advertising and sales promotion. Furthermore, Ottaviano (2001) has argued convincingly that a firm belonging to the modern sector could well use

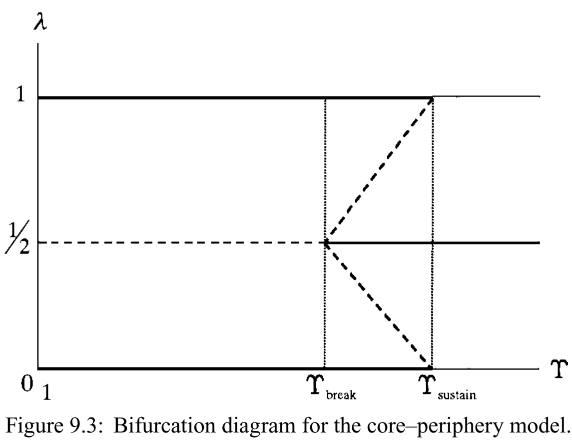


Figure 9.3: Bifurcation diagram for the core–periphery model.

unskilled labor to produce its output. Formally, this means that the production of any variety requires a fixed amount f of skilled labor and a marginal requirement $c = 1$ of unskilled labor. In this case, the marginal labor requirement is measured in terms of the numéraire. This vastly simplifies the analysis and allows for a more detailed treatment of the core–periphery model. Indeed, (9.12) becomes

$$\pi_r(i) = p_r(i)q_r(i) - w_r f - q_r(i) = [p_r(i) - 1]q_r(i) - w_r f \qquad r = A, B.$$

Because demands (9.4) are symmetric and isoelastic and the marginal cost is 1 instead of w_r, the equilibrium price is now the same across firms *and* regions:

$$p_A^* = p_B^* = \frac{1}{\rho}.$$

Thus, the delivered price Υ/ρ is also the same in both regions. Hence, the regional price indices are as follows:

$$P_r = \left[\lambda_r(H/f)(1/\rho)^{1-\sigma} + \lambda_s(H/f)(\Upsilon/\rho)^{1-\sigma}\right]^{1/(1-\sigma)} \qquad s \neq r$$

or

$$P_r = \frac{(H/f)^{1/(1-\sigma)}}{\rho}[\lambda_r + \lambda_s(\Upsilon)^{1-\sigma}]^{1/(1-\sigma)}. \tag{9.31}$$

The zero-profit output of a firm located in region $r = A, B$ is now

$$q_r^* = (\sigma - 1) f w_r,$$

which, unlike (9.15), varies with the region where the firm is located.

From (9.4), the demands in region $s = A, B$ for any variety produced in region $r = A, B$ are given by

$$q_s(r) = \frac{p_{rs}^{-\sigma}}{M_r p_{rs}^{1-\sigma} + M_s p_{ss}^{1-\sigma}} \mu Y_s$$

in which the regional incomes Y_r are

$$Y_r = L/2 + \lambda_r H w_r \qquad r = A, B.$$

Consequently, the market-clearing conditions for the differentiated product are given by

$$q_r^* = q_r(i) + \Upsilon q_s(i) \qquad s \neq r$$

or

$$(\sigma - 1)w_r = \frac{\rho\mu}{\lambda_r H + \lambda_s H \Upsilon^{1-\sigma}}\left(\frac{L}{2} + \lambda_r H w_r\right) + \frac{\rho\mu\Upsilon^{1-\sigma}}{\lambda_r H \Upsilon^{1-\sigma} + \lambda_s H}\left(\frac{L}{2} + \lambda_s H w_s\right),$$

which are linear in the nominal wages w_r and w_s.

Given the solutions to these two equations, (9.20) and (9.31) imply that

$$\Omega(\lambda) \equiv \frac{\omega_A(\lambda)}{\omega_B(\lambda)} = \frac{w_A^*/P_A^\mu}{w_B^*/P_B^\mu} = \left[\frac{\lambda + \Upsilon^{1-\sigma}(1-\lambda)}{(1-\lambda) + \Upsilon^{1-\sigma}\lambda}\right]^{\frac{\mu}{\sigma-1}}$$
$$\times \frac{2\Upsilon^{1-\sigma}(\sigma-1)\lambda + \left[(\sigma-\mu\rho-1) + \Upsilon^{2(1-\sigma)}(\sigma+\mu\rho-1)\right](1-\lambda)}{2\Upsilon^{1-\sigma}(\sigma-1)(1-\lambda) + \left[(\sigma-\mu\rho-1) + \Upsilon^{2(1-\sigma)}(\sigma+\mu\rho-1)\right]\lambda}.$$

Because

$$\Delta\omega(\lambda) = \omega_B(\lambda)[\Omega(\lambda) - 1]$$

we have

$$\omega(\lambda) \gtreqless 0 \quad \text{if and only if} \quad \Omega(\lambda) \gtreqless 1.$$

Thus, $\Omega(\lambda)$ allows us to obtain the indirect utility differential as an explicit function of the distribution of skilled workers, which is something that cannot be achieved in the original core–periphery model. As usual, because $\Omega(1/2) = 1$, $\lambda = 1/2$ is always a spatial equilibrium.

By studying $\Omega(\lambda)$ with respect to λ, one can show that the equilibrium configurations are similar to those obtained above. First, let us determine Υ_{break}, which is a hard task in the original model. Because $\Omega(1/2) = 1$, the symmetric configuration is stable for some Υ if and only if

$$\Omega'(1/2) < 0$$

for the corresponding value of Υ. Setting $\phi \equiv \Upsilon^{1-\sigma}$, we see that ϕ varies between 0 (prohibitive transportation costs) and 1 (zero transportation costs). It can readily be verified that

$$\Omega'(1/2) = 0$$

has two roots in ϕ:

$$\phi_1 = \frac{(\sigma-\mu-1)(\sigma-\mu)}{(\sigma+\mu-1)(\sigma+\mu)} \quad \text{and} \quad \phi_2 = 1$$

with $\phi_1 < \phi_2$. When $\mu \geq \sigma - 1$, $\phi_1 < 0$ so that the symmetric equilibrium is never stable. This means that this condition is comparable to the black hole condition $\mu \geq \rho \equiv (\sigma-1)/\sigma$ obtained in Section 9.2.3. When $\mu > \sigma - 1$, $\partial\Omega/\partial\lambda$ evaluated at $\lambda = 1/2$ and $\phi = 0$ is negative, and thus the symmetric configuration is stable for any $\phi < \phi_1$. In other words, the symmetric configuration is stable provided that Υ is larger than

$$\Upsilon_{\text{break}} = \left[\frac{(\sigma+\mu-1)(\sigma+\mu)}{(\sigma-\mu-1)(\sigma-\mu)}\right]^{1/(\sigma-1)} > 1, \tag{9.32}$$

which is the counterpart of (9.30). For $\Upsilon < \Upsilon_{\text{break}}$, the symmetric equilibrium ceases to be stable. Next, setting $\Omega(1) = 1$ yields

$$\frac{1+\mu/\sigma}{2}\Upsilon^{-(\sigma-1+\mu)} + \frac{1-\mu/\sigma}{2}\Upsilon^{\sigma-1-\mu} = 1, \tag{9.33}$$

which is the counterpart of (9.29). Solving this equation, we obtain the $\Upsilon_{\text{sustain}}$ such that the core–periphery structure is a stable equilibrium for any $\Upsilon \leq \Upsilon_{\text{sustain}}$.

9.3 STICKY LABOR AND REGIONAL SPECIALIZATION

In the core–periphery model, agglomeration is the outcome of a circular causation process in which more people concentrate within the same region because they love variety. However, if workers are sticky, no agglomeration can arise. Instead, each region specializes in the production of differentiated varieties on the basis of their initial endowments, and intraindustry trade occurs for all values of the transportation costs.

However, the agglomeration of industries is a pervasive phenomenon even when labor is immobile. An alternative way to explain the emergence of agglomeration is to recognize that the modern sector uses an array of differentiated intermediate goods. In this case, *the agglomeration of the modern sector in a particular region can occur because of the concentration of the intermediate industry in that region, and conversely*. In this section, we show how this process works in a context in which we combine elements belonging to the core–periphery model and to the one presented in Section 4.2.1.[17] The key issue in the approach followed here is how workers living in a given region allocate themselves between the different sectors of the regional economy assuming for simplicity that the intersectoral mobility is perfect. For each region, a given allocation of labor generates a certain wage through the labor market clearing condition. At the corresponding wages, firms choose to stay put or to delocate. In equilibrium, no firm of the modern or intermediate sector has an incentive to change location.

9.3.1 The Framework

The economy involves three sectors: the final, intermediate, and traditional ones. Because workers are spatially immobile, love for variety is no longer an agglomeration force, and thus it is convenient to assume that the output of the modern sector is homogeneous. Preferences are identical across all workers and described by the utility (9.1):

$$u = Q^{\mu}T^{1-\mu}/\mu^{\mu}(1-\mu)^{1-\mu} \qquad 0 < \mu < 1,$$

where Q now stands for the consumption of a homogeneous good produced by the modern sector, and T is the consumption of the traditional sector's output.

Workers being immobile, we may consider a single type of labor. Because the output is homogeneous, the $\mathbb{M}$-sector is assumed to operate under constant returns to scale and perfect competition. The $\mathbb{M}$-good is produced according to the production function:

$$X^{\mathbb{M}} = l^{1-\alpha} I^{\alpha} \qquad 0 < \alpha < 1, \tag{9.34}$$

where $X^{\mathbb{M}}$ is the output of the $\mathbb{M}$-sector, l the amount of labor, and I represents an index of the consumption of the intermediate varieties defined by

$$I = \left\{ \int_0^M [q(i)]^{\rho} di \right\}^{1/\rho} \qquad 0 < \rho < 1 \tag{9.35}$$

in which $q(i)$ is the quantity of the intermediate good i and M the number of intermediate goods. As usual, a smaller ρ means a more differentiated set of intermediate varieties.

By contrast, the intermediate sector exhibits increasing returns and operates under monopolistic competition (as in Section 4.2.1). Each variety of the $\mathbb{I}$-sector is produced according to the same technology such that the production of the quantity $q(i)$ requires $l(i)$ units of labor, which is again given by

$$l(i) = f + q(i), \tag{9.36}$$

where f is the fixed requirement of labor.

Finally, as in Section 9.2, the technology in the $\mathbb{T}$-sector is such that one unit of output requires one unit of labor.

The demand functions for the two consumption goods are as follows:

$$T = (1 - \mu)Y/p^{\mathbb{T}} \tag{9.37}$$

$$Q = \mu Y/p^{\mathbb{M}}, \tag{9.38}$$

where $p^{\mathbb{M}}$ is the price of the $\mathbb{M}$-good.

The price index for the $\mathbb{I}$-sector is as given by (9.5):

$$P \equiv \left[\int_0^M p(i)^{-(\sigma-1)} di \right]^{-1/(\sigma-1)}, \tag{9.39}$$

where $\sigma \equiv 1/(1 - \rho)$ and $p(i)$ is the price of the intermediate good i. Given the wage rate w, the unit production cost in the $\mathbb{M}$-sector is

$$c^{\mathbb{M}} = \alpha^{-\alpha}(1 - \alpha)^{-(1-\alpha)} w^{1-\alpha} P^{\alpha}, \tag{9.40}$$

whereas the input demands of the $\mathbb{M}$-sector corresponding to output $X^{\mathbb{M}}$ are

$$L^{\mathbb{M}} = (1 - \alpha) c^{\mathbb{M}} X^{\mathbb{M}} w^{-1} \tag{9.41}$$

$$q(i) = \alpha c^{\mathbb{M}} X^{\mathbb{M}} p(i)^{-\sigma} P^{\sigma-1} \qquad i \in [0, M]. \tag{9.42}$$

Consider an economy with two regions A and B, each endowed with $L > 0$ workers. As in the foregoing section, we assume that the $\mathbb{T}$-good can be shipped costlessly from one region to another; this good is used as the numéraire ($p^{\mathbb{T}} = 1$). The output of the $\mathbb{M}$-sector ($\mathbb{I}$-sector) is shipped from one region to the other at a positive cost according to the iceberg cost $\Upsilon^{\mathbb{M}} > 1$ ($\Upsilon^{\mathbb{T}} > 1$). Let $L_r^{\mathbb{M}}$, $L_r^{\mathbb{I}}$, and $L_r^{\mathbb{T}}$ be the mass of workers living in region r ($= A, B$) and working in the modern, intermediate, and traditional sectors, respectively.

Using the same argument as in Section 9.2.2, we see that the common equilibrium price of the intermediate varieties produced in region $r = A, B$ is

$$p_r^{\mathbb{I}} = \frac{w_r}{\rho},$$

and thus the output of an $\mathbb{I}$-sector firm under zero profit is still given by $(\sigma - 1)f$ and its labor requirement by σf.

The price index for the $\mathbb{I}$-goods in region $r = A, B$ can be shown to be equal to

$$P_r = k_1 \left[L_r^{\mathbb{I}}(w_r)^{-(\sigma-1)} + L_s^{\mathbb{I}}(w_s \Upsilon^{\mathbb{I}})^{-(\sigma-1)}\right]^{-1/(\sigma-1)} \qquad s \neq r, \tag{9.43}$$

where

$$k_1 \equiv \rho^{-1}(\sigma f)^{1/(\sigma-1)}.$$

The common output of a $\mathbb{I}$-firm located in r is

$$q_r = \alpha \left(\frac{w_r}{\rho}\right)^{-\sigma} \left[c_r^{\mathbb{M}} X_r^{\mathbb{M}} P_r^{\sigma-1} + c_s^{\mathbb{M}} X_s^{\mathbb{M}} (\Upsilon^{\mathbb{I}})^{-(\sigma-1)} P_s^{\sigma-1}\right] \qquad s \neq r \tag{9.44}$$

where $X_r^{\mathbb{M}}$ is the production of the $\mathbb{M}$-good in region $r = A, B$.

9.3.2 Agglomeration of the Intermediate and Final Sectors

Suppose that both the final and intermediate industries are concentrated in one region, say region A, so that $L_B^{\mathbb{M}} = L_B^{\mathbb{I}} = 0$ and $L_B^{\mathbb{T}} = L$. Hence, region A exports the $\mathbb{M}$-good and region B the $\mathbb{T}$-good. Assume also that workers residing in region A work either for the intermediate or for the modern sector so that $L_A^{\mathbb{T}} = 0$. Then, it follows that $w_A^* \geq w_B^* = 1$.

The corresponding regional price indices are obtained from (9.43) as follows:

$$\begin{aligned} P_A &= k_1 \left[L_A^{\mathbb{I}}(w_A^*)^{-(\sigma-1)}\right]^{-1/(\sigma-1)} \\ &= k_1 \frac{\mu}{1-\mu} \left(L_A^{\mathbb{I}}\right)^{-1/(\sigma-1)} \end{aligned} \tag{9.45}$$

$$P_B = P_A \Upsilon^{\mathbb{I}}. \tag{9.46}$$

Consider first the equilibrium of the $\mathbb{T}$-good market. Because $Y_A = Lw_A^*$ and $Y_B = L$, regional demands are respectively given by

$$T_A = (1-\mu)Lw_A^* \quad \text{and} \quad T_B = (1-\mu)L.$$

Because $X_A^{\mathbb{T}} = 0$ and $X_B^{\mathbb{T}} = L$, the equality of supply and demand implies that

$$w_A^* = \frac{\mu}{1-\mu}. \tag{9.47}$$

For the traditional sector to be unprofitable in region A, it follows that $w_A^* \geq 1$, which holds if and only if

$$\mu \geq 1/2. \tag{9.48}$$

Because we have

$$p_A^{\mathbb{M}} = c_A^{\mathbb{M}} \quad \text{and} \quad p_B^{\mathbb{M}} = p_A^{\mathbb{M}}\Upsilon^{\mathbb{M}} = c_A^{\mathbb{M}}\Upsilon^{\mathbb{M}},$$

using (9.38), the regional demands for the $\mathbb{M}$-good are given by

$$Q_A = \mu Lw_A^*/c_A^{\mathbb{M}} \quad \text{and} \quad Q_B = \mu L/c_A^{\mathbb{M}}\Upsilon^{\mathbb{M}}.$$

Because the $\mathbb{M}$-good is exported to B, the equilibrium of the $\mathbb{M}$-good market implies that the total production $X_A^{\mathbb{M}}$ is such that

$$\begin{aligned} X_A^{\mathbb{M}} &= Q_A + Q_B\Upsilon^{\mathbb{M}} \\ &= \frac{\mu}{1-\mu}\frac{L}{c_A^{\mathbb{M}}} \end{aligned}$$

from (9.47), or

$$X_A^{\mathbb{M}}c_A^{\mathbb{M}} = \frac{\mu}{1-\mu}L. \tag{9.49}$$

Producing the $\mathbb{M}$-good in region B is never profitable if and only if

$$c_B^{\mathbb{M}} \geq p_B^{\mathbb{M}} = c_A^{\mathbb{M}}\Upsilon^{\mathbb{M}},$$

which amounts to $c_B^{\mathbb{M}}/c_A^{\mathbb{M}} \geq \Upsilon^{\mathbb{M}}$. With (9.40), this holds if and only if

$$\Upsilon^{\mathbb{I}} \geq \left(\frac{\mu}{1-\mu}\right)^{(1-\alpha)/\alpha} (\Upsilon^{\mathbb{M}})^{1/\alpha}.$$

It remains to consider the $\mathbb{I}$-sector. Given (9.44), (9.49), and $X_B = 0$, we have

$$\begin{aligned} q_A &= \alpha\left(\frac{w_A}{\rho}\right)^{-\sigma}\left[c_A^{\mathbb{M}}X_A^{\mathbb{M}}P_A^{\sigma-1}\right] \\ &= \alpha\rho^{\sigma}(k_1)^{\sigma-1}L\left(L_A^{\mathbb{I}}\right)^{-1} \end{aligned}$$

$$\begin{aligned} q_B &= \alpha\left(\frac{1}{\rho}\right)^{-\sigma}\left[c_A^{\mathbb{M}}X_A^{\mathbb{M}}(\Upsilon^{\mathbb{I}})^{-(\sigma-1)}P_A^{\sigma-1}\right] \\ &= \rho^{\sigma}k_1^{\sigma-1}\left(\frac{\mu}{1-\mu}\right)^{\sigma}(\Upsilon^{\mathbb{I}})^{-(\sigma-1)}. \end{aligned}$$

The first equilibrium condition is $q_A = q^* = (\sigma - 1)f$, which yields

$$L_A^{\mathbb{I}} = \alpha L.$$

The second equilibrium condition, that is, the nonprofitability of region B for $\mathbb{I}$-firms, means $q_B \leq q^*$, which is equivalent to the condition

$$\Upsilon^{\mathbb{I}} \geq \left(\frac{\mu}{1-\mu}\right)^{1/\rho}.$$

To sum up, we have shown the following:

Proposition 9.2 *Assume $\mu \geq 1/2$. Then, the agglomeration of the intermediate and final sectors into the same region is an equilibrium if and only if the following two conditions are satisfied:*

$$\Upsilon^{\mathbb{I}} \geq \left(\frac{\mu}{1-\mu}\right)^{(1-\alpha)/\alpha} (\Upsilon^{\mathbb{M}})^{1/\alpha} \tag{9.50}$$

$$\Upsilon^{\mathbb{I}} \geq \left(\frac{\mu}{1-\mu}\right)^{1/\rho}. \tag{9.51}$$

Hence, when the transport cost of the intermediate goods is high relative to the transport cost of the final good, there is complete regional specialization because both the final and intermediate sectors are entirely concentrated in one region, whereas the traditional sector operates only in the other region. Condition (9.50) becomes less stringent as the transport cost of the final good declines. In addition, the transport cost of the intermediate goods must also exceed some threshold value (9.51), for $\mu/(1-\mu) \geq 1$. Clearly, this threshold rises when the intermediate goods are more differentiated. The domains of $(\Upsilon^{\mathbb{I}}, \Upsilon^{\mathbb{M}})$ sustaining the core–periphery structure are represented by the shaded areas in Figure 9.4.

Condition (9.50) means that the $\mathbb{M}$-sector does not find it profitable to start operating in region B because importing the intermediate goods from A turns out to be costly owing to the high transport costs of these goods, whereas exporting its output from A to B is less costly because of the relatively low value of $\Upsilon^{\mathbb{M}}$. Condition (9.51) means that no firm of the intermediate sector wants to set up in region B because it has to export all its production to region A at a high transport cost. It should be stressed that both sectors are trapped within the same region even when shipping the final good becomes cheaper and cheaper ($\Upsilon^{\mathbb{M}}$ approaches 1).

To break such a trap, the transport costs of the intermediate goods must fall below some critical value. This is not necessarily easy to implement when the provision of specific intermediate goods requires face-to-face contacts such as for highly differentiated services (in which case $\Upsilon^{\mathbb{I}}$ is high).

As long as (9.50) and (9.51) hold, μ can rise. Because $w_A^* = \mu/(1-\mu)$ and $w_B^* = 1$, this rise generates a widening wage gap between the core region

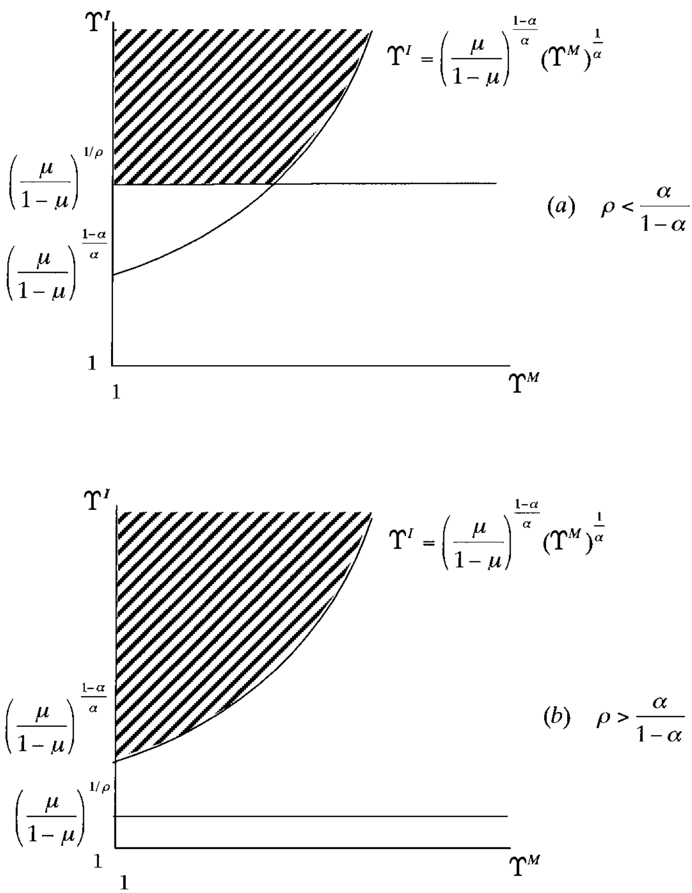


Figure 9.4: Transport cost ranges for sustaining the core–periphery structure.

and the periphery. It can readily be verified that the real wage gap, in turn, becomes even larger. This agrees with the observation that (9.50) becomes less and less stringent as the role of the intermediate goods plays a growing role in the economy (α rises). However, the modern sector will eventually decentralize some of its activities in the periphery as its share in consumption increases.

When transportation costs for the intermediate sector decrease enough for (9.50) not to hold anymore, whereas (9.51) is still valid, one expects the $\mathbb{I}$-firms to remain concentrated in region A, and some share of the final sector now operates in region B. In this case, the intermediate goods needed by the final sector in region B are imported from region A. Finally, when shipping the intermediate goods becomes very inexpensive ($\Upsilon^{\mathbb{I}}$ approaches 1), it is reasonable to expect the symmetric configuration to be the only stable equilibrium. These issues are left for further investigation.

9.4 A LINEAR MODEL OF CORE–PERIPHERY: DISCRIMINATORY PRICING AND WELFARE

The setting considered in this section is very similar to the one used in Section 9.2. However, there are two major differences. First, the output of the $\mathbb{M}$-sector is traded at a cost of t units of the numéraire per unit shipped between regions. This characteristic agrees more with reality as well as with location theory than the iceberg technology does. In particular, as will be demonstrated, this allows us to study and compare various spatial price policies such as discriminatory and mill pricing. Second, preferences are given by a quasi-linear utility encapsulating a quadratic subutility instead of a Cobb–Douglas preference on the homogeneous and differentiated goods with CES subutility.[18] These two specifications correspond to two rather extreme cases: the former assumes an infinite elasticity of substitution between the differentiated product and the numéraire, the latter a unit elasticity. Moreover, firms' demands are linear instead of iso-elastic.

Specifically, we use the model described in Section 8.4, which is now embedded in a general equilibrium framework. In particular, preferences are identical across all workers and described by (8.26) with individual demand $q(i)$ for variety i given by

$$q(i) = a - (b + dM)\,p(i) + dP, \tag{9.52}$$

where $p(i)$ is the price of variety $i \in [0, M]$ and P/M the price index in the modern sector. Hence, each worker has the following quasi-linear indirect utility function:

$$v(y; p(i),\ i \in [0, M]) = \frac{a^2 M}{2b} - a\int_0^M p(i)di + \frac{b + dM}{2}\int_0^M [p(i)]^2 di - \frac{d}{2}\left[\int_0^M p(i)di\right]^2 + y + \bar{z}, \tag{9.53}$$

where y is the worker's labor income and $\bar{z}$ his initial endowment in the numéraire. It is assumed that each worker's initial endowment $\bar{z}$ in the numéraire is large enough for his residual consumption of the numéraire to be strictly positive in equilibrium.[19] The technologies are the same as in Section 9.2 but, for simplicity, c is set equal to zero in (9.7).[20]

As stated in the introduction, we focus on *discriminatory pricing*. Recall that, in this case, each firm sets a delivered price specific to the market in which its variety is sold; hence, markets are segmented. In addition, because the iceberg cost enters the profit function multiplicatively in Krugman's model, demands have the same elasticity across locations, and thus both mill and discriminatory pricing policies yield the same equilibrium prices and outputs. Few people noticed the equivalence between the two policies under the iceberg technology in the spatialized S–D–S model (Hsu 1979; Greenhut et al. 1987, chap. 2). This equivalence no longer holds in the present model.

9.4.1 Equilibrium Prices

Suppose that each firm sets a (delivered) price specific to each region. Hence, the profit function of a representative firm located in region $r = A, B$ is as follows:

$$\pi_r = p_{rr} q_{rr}(p_{rr})(L/2 + \lambda_r H) + (p_{rs} - t) q_{rs}(p_{rs})(L/2 + \lambda_s H) - f w_r \qquad s \neq r, \tag{9.54}$$

where $\lambda \equiv \lambda_A$ and $\lambda_B = 1 - \lambda$.

To illustrate the type of interaction that characterizes this model of monopolistic competition, we describe how the equilibrium prices are determined. Each firm i in region r maximizes its profit π_r, assuming accurately that its price choice has no impact on the regional price indices

$$P_r \equiv \int_0^{M_r} p_{rr}(i) di + \int_0^{M_s} p_{sr}(i) di \qquad s \neq r.$$

Because, by symmetry, the prices selected by the firms located within the same region are identical, the resulting prices are denoted by $p^*_{rr}(P_r)$ and $p^*_{rs}(P_s)$. Clearly, it follows that

$$M_r p^*_{rr}(P_r) + M_s p^*_{sr}(P_r) = P_r \qquad s \neq r.$$

By solving the first-order conditions for profit maximization with respect to prices, it can readily be verified that the equilibrium prices are as follows:

$$p^*_{rr} = \frac{1}{2} \frac{2a + dt(1 - \lambda_r)M}{2b + dM} \qquad s \neq r \tag{9.55}$$

$$p^*_{rs} = p^*_{ss} + \frac{t}{2} \qquad s \neq r. \tag{9.56}$$

Clearly, these prices depend directly on the firms' (or workers') distribution across regions. More precisely, p^*_{rr} decreases with the mass of firms in region r, $\lambda_r M$, and increases with the degree of product differentiation, which decreases with d. These results agree with what we know from models of product differentiation (see, e.g., Anderson et al. 1992, chap.7). In addition, these prices also show that arbitrage is never profitable because $p^*_{rs} - p^*_{rr} < \tau$.

It is easy to check that the equilibrium operating profits earned in each market by a firm established in $r = A, B$ are as follows:

$$\pi^*_{rr} = (b + dM)(p^*_{rr})^2 \, (L/2 + \lambda_r H)$$

$$\pi^*_{rs} = (b + dM)(p^*_{rs} - t)^2 \, (L/2 + \lambda_s H),$$

where π^*_{rk} denotes the profits earned from selling in region $k = r, s$.

Increasing λ_r has two opposite effects on π^*_{rr}. First, as λ_r rises, both the equilibrium price (9.55) and the corresponding quantity of each variety bought by each consumer living in region r, fall. Second, the total population of consumers residing in this region is now larger, and thus the profits made by a firm located in r on local sales may increase. What is at work here is a global demand effect resulting from the increase in the local population that may compensate firms for the adverse price effect as well as for the decrease in each worker's individual demand.

Deducting t from (9.55) and (9.56), we see that firms' prices net of transport costs are positive regardless of the workers' distribution if and only if

$$t < t_{\text{trade}} \equiv \frac{2af}{2bf + dH} \equiv \frac{2\alpha f}{2f + \delta H/(\beta - \delta)}, \tag{9.57}$$

which depends only upon the primitives $(L, H, \alpha, \beta, \delta, f)$ once a, b, d, and M are replaced by their values. More generally, it can readily be verified that

$$\frac{d\tau_{\text{trade}}}{df} > 0 \qquad \frac{d\tau_{\text{trade}}}{d\delta} < 0;$$

thus, trade is more likely the higher are the intensity of increasing returns and the degree of product differentiation. The condition (9.57) must also hold for consumers in s to buy from firms in r, that is, for the individual demands (9.52) in each region evaluated at the equilibrium prices to be positive for all λ. Observe that there is no trade when there are no increasing returns ($f = 0$). In this case, each region supplies all the varieties and is in autarchy – a result that agrees with what we saw in Chapter 2.

The consumer surplus $C_r(\lambda)$ of each worker in region r associated with the equilibrium prices (9.55) and (9.56) is then as follows:

$$C_r(\lambda) = \frac{a^2 H}{2bf} - \frac{aH}{f}(\lambda_r p^*_{rr} + \lambda_s p^*_{sr}) + \frac{(bf + dH)H}{2f^2}\left[\lambda_r (p^*_{rr})^2 + \lambda_s (p^*_{sr})^2\right]$$
$$- \frac{dH^2}{2f^2}(\lambda_r p^*_{rr} + \lambda_s p^*_{sr}).$$

Differentiating this expression twice with respect to λ shows that $C_r(\lambda)$ is concave. Furthermore, (9.57) implies that $C_r(\lambda)$ is always increasing over the interval [0, 1].

Labor market clearing implies that the mass M_r of firms belonging to the $\mathbb{M}$-sector in region r is

$$M_r = \lambda_r H/f. \tag{9.58}$$

Consequently, the total mass of firms in the economy is fixed and equal to $M = M_A + M_B = H/f$. As in Section 9.2, the number of firms is constant. In addition, (9.58) shows that the region with the larger number of workers is also the region accommodating the larger number of firms.

Entry and exit are free; therefore, profits are zero in equilibrium. Hence, (9.58) implies that any change in the population of workers located in one region must be accompanied by a corresponding change in the number of firms. The equilibrium wage rates w_r^* of the skilled are obtained from the zero profit condition evaluated at the equilibrium prices, whereas the wage rate of the unskilled is equal to their marginal product, that is, 1.

More precisely, the equilibrium wage prevailing in region $r = A, B$ may be obtained from (9.54) by computing $w_r^*(\lambda) = (\pi_{rr}^* + \pi_{rs}^*)/f$. This yields the following expression:

$$w_r^*(\lambda) = \frac{bf + dH}{4(2bf + dH)^2 f^2}\left\{(2af + \tau dH\lambda_s)^2\left(\frac{L}{2} + \lambda_r H\right)\right.$$
$$\left. + (2af - 2\tau bf - \tau dH\lambda_s)^2\left(\frac{L}{2} + \lambda_s H\right)\right\}, \qquad (9.59)$$

which, after simplification, turns out to be quadratic in λ. Standard but cumbersome investigations reveal that $w_r^*(\lambda)$ is concave and increasing (convex and decreasing) in λ when f is large (small) as well as when t, d, H, and L are small (large). This implies that both $C_r(\lambda)$ and $w_r^*(\lambda)$ increase with λ when t is small, whereas they go in opposite directions when t is large. This provides useful information for the study of the agglomeration process.

9.4.2 Agglomeration versus Dispersion

The indirect utility differential is obtained by plugging the equilibrium prices (9.55) and (9.56) and the equilibrium wages (9.59) into (9.53):

$$\Delta v(\lambda) \equiv v_A(\lambda) - v_B(\lambda) \equiv C_A(\lambda) - C_B(\lambda) + w_A^*(\lambda) - w_B^*(\lambda)$$
$$= C_M t(t^* - t)(\lambda - 1/2), \qquad (9.60)$$

where

$$C_M \equiv [2bf(3bf + 3dH + dL) + d^2H(L + H)]\frac{H(bf + dH)}{2f^2(2bf + dH)^2} > 0$$

and

$$t^* \equiv \frac{4af(3bf + 2dH)}{2bf(3bf + 3dH + dL) + d^2H(L + H)}, \qquad (9.61)$$

which can also be restated in terms of the primitives of the economy. The stability of an equilibrium is studied with respect to (9.25).

Observe that $\Delta v(\lambda)$ is linear in λ. It follows immediately from (9.60) that $\lambda = 1/2$ is always an equilibrium. Because $C_M > 0$, for $\lambda \neq 1/2$, the indirect utility differential always has the same sign as $\lambda - 1/2$ if and only if $t < t^*$; if $t > t^*$, it has the opposite sign. In particular, when there are no increasing returns in the manufacturing sector ($f = 0$), the coefficient of $(\lambda - 1/2)$ is always negative because $t^* = 0$; thus, symmetry is the only (stable) equilibrium.[21] This shows once more the importance of increasing returns for the possible emergence of an agglomeration. The same holds for product differentiation because t^* becomes arbitrarily small when varieties become less and less differentiated ($d \to \infty$).[22]

It remains to determine when t^* is lower than t_{trade}, as given by (9.57). This is so if and only if

$$L/H > \frac{6b^2 f^2 + 8bdfH + 3d^2 H^2}{dH(2bf + dH)} > 3, \tag{9.62}$$

where the second inequality holds because $b/d = \beta/\delta - 1 \in (0, \infty)$. The inequality (9.62) means that the population of unskilled is large relative to the population of skilled. When (9.62) does not hold, the coefficient of $(\lambda - 1/2)$ in (9.60) is always positive for all $t < t_{\text{trade}}$. In this case, the advantages of having a large home market always dominate the disadvantages incurred while supplying a distant periphery. The reverse of (9.62) plays a role similar to the black hole condition described in Section 9.2.3.

More interesting is the case when condition (9.62) holds. Although the size of the industrial sector is captured here through the relative population size L/H and not through its share in consumption, the intuition is similar: the ratio L/H must be sufficiently large for the economy to display different types of equilibria according to the value of t. This result does not depend on the expenditure share on the manufacturing sector because general equilibrium income effects are absent: either small or large sectors in terms of expenditure share are agglomerated when t is small enough.

When $t > t^*$, it is straightforward to see that the symmetric configuration is the only stable equilibrium. In contrast, when $t < t^*$ the symmetric equilibrium becomes unstable and workers agglomerate in region A (B) provided that the initial fraction of workers residing in this region exceeds $1/2$. In other words, *agglomeration arises when transportation costs are sufficiently low*, as in Section 9.2.3 and for similar reasons. In addition, because the incentives to move keep rising with the size of the agglomeration, once made, no worker regrets his choice to live in the core region.

Proposition 9.3 *Consider a two-region economy with segmented markets.*

1. *When condition (9.62) does not hold, the core–periphery structure is the only stable equilibrium under trade.*
2. *When (9.62) is satisfied, for any $t > t^*$ the symmetric configuration is the only stable equilibrium with trade; for any $t < t^*$ the core–periphery pattern is the unique stable equilibrium; for $t = t^*$ any configuration is an equilibrium.*

As t decreases from some value larger than t^* but smaller than t_{trade}, we move from a stable dispersed equilibrium with two symmetric regions to a market situation in which any stable equilibrium involves a core–periphery pattern. Note that here the break point and the sustain point are the same and that history alone matters for the selection of the agglomerated outcome; this follows because (9.60) is linear in λ.

Looking at the threshold value t^*, as given by (9.61), we first observe that it increases when the intensity of preference for the $\mathbb{M}$-good relative to the $\mathbb{T}$-good (as measured by a higher α or a) increases. Second, t^* increases with the degree of product differentiation (δ or d falls) when (9.62) holds.[23] Third, higher fixed costs lead to a smaller mass of firms and varieties. Still, it can readily be verified that t^* also increases when increasing returns become stronger (f rises) when (9.62) holds. In other words, the agglomeration of the modern sector is more likely the stronger are the increasing returns at the firm's level. Last, t^* increases when the mass of unskilled (L) decreases because the dispersion force is weaker.

All these results are similar to those obtained in Section 9.2, which suggests some robustness in the reasons for the core–periphery structure because the two models considered differ significantly.

Because the skilled are mobile, they always reach the same utility level, and thus agglomeration does not entail any redistribution within this group. However, agglomeration gives rise to redistributive effects inside the population of unskilled. Since, as seen above, the surplus C_A increases with λ, any increase from $\lambda = 1/2$ makes the unskilled in A better off and, by symmetry, the unskilled in B worse off.[24] Furthermore, given that C_A is concave in λ, the marginal gain of the former is lower than the marginal loss of the latter: agglomeration hurts the unskilled in the periphery more than it helps those living in the core region.

9.4.3 A Diagrammatic Analysis

It is possible to convey the economic intuition behind Proposition 9.3 through a simple graphical analysis. Figure 9.5 depicts the aggregate inverse demand

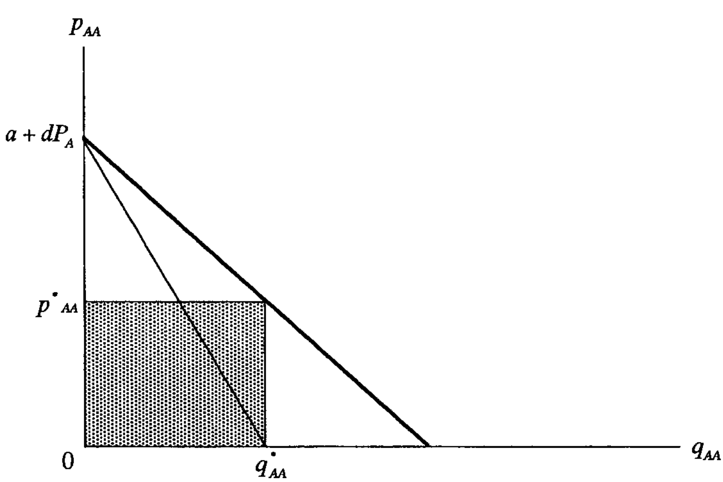


Figure 9.5: The impact of an increase in the number of firms on demand.

in, say, region A for a typical local variety; for simplicity, the unit of the differentiated product is chosen such that $b + dM = 1$:

$$p_{AA} = a + dP_A(M_A, t) - \frac{q_{AA}}{L/2 + fM_A}. \tag{9.63}$$

Because $p_{BA} > p_{AA}$ and the total number of firms is fixed by skilled labor market clearing, the price index P_A is a decreasing function of M_A at a rate that increases with t:

$$\frac{\partial P_A(M_A, t)}{\partial M_A} < 0 \quad \text{and} \quad \left| \frac{\partial^2 P_A(M_A, t)}{\partial M_A \partial t} \right| > 0. \tag{9.64}$$

The horizontal and vertical intercepts of (9.63) are, respectively, $[a + dP_A(M_A, t)]$ times $(L/2 + fM_A)$ and $[a + dP_A(M_A, t)]$. The equilibrium values of q_{AA} and p_{AA} are shown as q^*_{AA} and p^*_{AA}. They are found by setting marginal revenue equal to marginal cost (here zero). The equilibrium operating profits are given by the shaded rectangle and accrue to the skilled workers, whereas, as usual, the triangle above this rectangle represents the consumer surplus enjoyed by both skilled and unskilled workers.

Figure 9.5 is a powerful learning device to understand the forces at work in the model. To see why, start from an initial situation in which regions are identical ($M_A = M_B$). Suppose that some firms move from the foreign to the home region so that M_A rises and M_B falls. For these firms to stay in region A, their operating profits should not decrease. An increase in M_A has two opposite effects on operating profits. First, as new firms enter region A, the price index $P_A(M_A, t)$ decreases. Ceteris paribus, this would shift the inverse demand (9.63) toward the origin of the axes and operating profits would shrink. This effect is due to increased competition in the corresponding market and occurs because fewer firms now face transportation costs when supplying their home market. But this negative competition effect is not the only one. For some firms to move to the home region, some skilled workers have to follow ($M_A = \lambda H/f$). This means that, as M_A increases, λ also goes up so that the market of region A expands. Ceteris paribus, the horizontal intercept of the inverse demand would move away from the origin and profits would expand. This is a positive demand effect induced by the linkage between the locations of firms and skilled workers' expenditures.

Because the two effects oppose each other, the net result is a priori ambiguous. But we can say more than that. In particular, we can assess which effect prevails, depending on parameter values. Start with the competition effect that goes through $[a + dP_A(M_A, t)]$. This effect is strong if d is large, that is, if varieties are close substitutes. It is also strong if $|\partial P_A(M_A, t)/\partial M_A|$ is large. As shown in (9.64), this happens if t is large because, when obstacles to trade are high, competition from the other region is weak and home firms care more about their competitors being close rather than distant. As for the demand effect, it will be strong if f is large, because each new firm brings along many skilled workers, and if L is small because the skilled have a relatively large impact on the local market size.

We can therefore conclude that the demand effect dominates the competition effect; consequently, agglomeration occurs when goods are very differentiated (d small), increasing returns are strong (f large), and transportation costs are low (t small). Under such circumstances, the entry of new firms in one region raises the operating profits of all firms. Higher profits would attract more firms, generating circular causation among firms and workers. Agglomeration is then sustainable as a spatial equilibrium.

9.4.4 A Welfare Analysis of the Core–Periphery Structure

We now wish to determine whether or not agglomeration is efficient, and if so, why and when (Ottaviano and Thisse 2001). A quasi-linear utility allows us to measure global efficiency by using the sum of individual utilities.

Different distortions and external effects are at work in the core–periphery model that suggest a significant discrepancy between equilibrium and optimum. Besides the standard distortion caused by firms' not pricing at marginal cost, there are several pecuniary externalities and, because our economy is imperfectly competitive, they matter for the level of welfare. In particular, skilled workers impose a pecuniary externality on the workers of the traditional sector. Indeed, their move affects the intensity of local competition in both the product and labor markets. In addition, when skilled workers move from one region to the other, they do not account for the impact of their migration on the other skilled. Their move affects not only the intensity of competition but also the level of demand inside each region, and, therefore, their wages. Finally, recall that there is no over- or under-entry effect. Indeed, the mass of firms is the same in equilibrium and at the optimum because it is determined by the technology and equal to H/f.

Let us first determine the optimum pattern.[25] For that, we assume that the planner is able to assign any mass of skilled workers (or, equivalently, of firms belonging to the modern sector) to a specific region and to use lump-sum transfers from all workers to pay for the loss firms incur while pricing at marginal cost. Since operating profits are just equal to the wage bill, the total welfare W in our two-region economy is given by the consumer surplus augmented by operating profits. The planner then chooses λ to maximize the total welfare W in our two-region economy given by (recall that individual utilities are quasi-linear)

$$W = \frac{L}{2}C_A + \lambda H(C_A + w_A) + \frac{L}{2}C_B + (1-\lambda)H(C_B + w_B) + L + \text{constant}, \tag{9.65}$$

where C_r is the consumer surplus in region $r = A, B$ in which all prices have been set equal to marginal cost:

$$p^0_{rr} = 0 \quad \text{and} \quad p^0_{rs} = t \quad s \neq r.$$

This implies that operating profits are zero; hence, $w^0_r(\lambda) = 0$ for every λ so that firms do not incur any loss. Consequently, (9.65) becomes

$$W(\lambda) = C_0 t(t^0 - t)\lambda(\lambda - 1) + \text{constant}, \tag{9.66}$$

where

$$C_0 \equiv \frac{H^2}{2f^2}[2bf + d(H + L)]$$

and

$$t^0 \equiv \frac{4af}{2bf + d(H + L)},$$

which, again, does not depend on M.

The function (9.66) is strictly concave in λ if $t > t^0$ and strictly convex if $t < t^0$. Furthermore, because the coefficients of λ^2 and of λ are the same (up to their sign), this expression always has an interior extremum at $\lambda = 1/2$. As a result, the optimal choice of the planner is determined by the sign of the coefficient of λ^2, that is, by the value of t with respect to t^0.

Hence we have the following:

Proposition 9.4 *Consider a two-region economy with segmented markets. If $t > t^0$, then the symmetric configuration is the first best optimum; if $t < t^0$, any agglomerated configuration is an optimum; if $t = t^0$, the spatial configuration does not affect the welfare level.*

As expected, it is socially desirable to agglomerate the modern sector into a single region once transportation costs are low, increasing returns in the modern sector are strong enough, the output of the modern sector is differentiated enough, or all three conditions obtain. The higher the degree of increasing returns, the higher is the critical value. In particular, agglomeration is never efficient ($t^0 = 0$) when there are no scale economies in production ($f = 0$) or when varieties are good substitutes ($d \to \infty$). In this case, both the market equilibrium and the optimum involve two self-sufficient regions that each produce all the varieties.

We now assume that lump sum transfers are not available to the planner, who is only able to assign locations to the skilled workers. In such a context, the social welfare function is still given by (9.65) but is now evaluated at the equilibrium prices (9.55) and (9.56) and wages (9.59). This leads to

$$W_S = C_S t(t^S - t)\lambda(\lambda - 1) + \text{constant},$$

where

$$C_S \equiv \frac{H^2(bf + dH)}{8f^2(2bf + dH)^2}[8bf(3bf + 2dH + dL) + 3d^2H(L + H)]$$

and

$$t^S \equiv \frac{16af(3bf + dH)}{8bf(3bf + 2dH + dL) + 3d^2H(H + L)}. \tag{9.67}$$

The choice of the planner is similar to that described in the first best case except that the threshold value of t is now given by t^S.

It can readily be verified that $t^* > 0$ (where t^* is given by (9.61)) implies $t^S > 0$. Hence we have the following:

Proposition 9.5 *Consider a two-region economy with segmented markets and assume that (9.62) does not hold. If $t > t^S$, then the symmetric configuration is the second best optimum; if $t < t^S$, any agglomerated configuration is a second best optimum; if $t = t^S$, the spatial configuration does not affect the welfare level.*

Some simple calculations reveal that $t^0 < t^S < t^*$ when (9.62) does not hold. These inequalities reveal several important things. First, $t^0 < t^S$, namely, agglomeration is desirable for higher values of the transport cost in the second best. This is because the individual demand elasticity is much lower in the first best (marginal cost pricing) than in the second best (Nash equilibrium pricing); thus, regional price indices are less sensitive to a decrease in t. The fall in transport costs must be sufficiently large to make the agglomeration of the mobile workers socially desirable.[26]

Second, we also have $t^S < t^*$. This is because skilled workers do not internalize the negative external effects they impose on the unskilled, who always prefer dispersion. Hence, even though the skilled have incentives to move, these incentives do not reflect the social value of their move. This discrepancy is even stronger when we compare the first best outcome and the market equilibrium because we have just seen that marginal cost pricing is more favorable to dispersion. As a result, the market yields the core–periphery structure for a whole range ($t^0 < t < t^*$) of transportation cost values under which it is socially desirable to have a dispersed pattern of activities.

Accordingly, when transport costs are low ($t < t^0$) or high ($t > t^*$), the market yields the optimal location pattern; thus, no regional policy is required from the efficiency point of view, although equity considerations might still justify such a policy when agglomeration arises. On the contrary, for intermediate values of transport costs ($t^0 < t < t^*$), the market provides excessive agglomeration, thus justifying the need for an active regional policy to foster the dispersion of the modern sector from both the efficiency and equity standpoints.[27]

To decipher the nature of the various spatial external effects at work in the core–periphery model, it is useful to consider the problem in which the planner maximizes (9.65) with respect to λ, controlling for the inefficiencies generated by noncompetitive pricing in the modern sector. The first-order condition requires equating the sum of three terms to zero:

$$W_S' = \Delta v H + [(C_A' + w_A')\lambda + (C_B' + w_B')(1-\lambda)]H + (C_A' + C_B')\frac{L}{2}.$$

On the right-hand side, the first term is the indirect utility differential that an independent skilled mover takes into account when migrating. The second and third terms stand for the external effects that the mover imposes on, respectively, the other skilled and unskilled workers. From previous results, we know that these three terms equate zero at $\lambda = 1/2$ so that the dispersed configuration always satisfies the first-order conditions of the planner ($W_S' = 0$), of the skilled as a whole ($\Delta v + [(C_A' + w_A')\lambda + (C_B' + w_B')(1-\lambda)] = 0$), and of the unskilled as a whole ($C_A' + C_B' = 0$), as well as the indifference condition of an individual skilled worker ($\Delta v = 0$). Consequently, any discrepancy between the market outcome and the optimum must arise from the second-order condition:

$$\begin{aligned} W_S'' &= 2[(C_A' + w_A') - (C_B' + w_B')]H \\ &\quad + [(C_A'' + w_A'')\lambda + (C_B'' + w_B'')(1-\lambda)]H + (C_A'' + C_B'')\frac{L}{2}, \end{aligned} \tag{9.68}$$

where we have used the fact that the derivative of Δv with respect to λ is equal to $(C_A' + w_A') - (C_B' + w_B')$. Condition (9.68) can be rearranged to give

$$\begin{aligned} W_S'' &= 2[(C_A' + w_A') - (C_B' + w_B')]H \\ &\quad + (C_A'' + C_B'')\frac{H+L}{2} + (w_A'' + w_B'')\frac{H}{2}, \end{aligned} \tag{9.69}$$

where the first term on the right-hand side is positive whenever agglomeration is the stable outcome, the second term on the right-hand side is always negative, and the third is positive (negative) whenever transportation costs are above (below) the value:

$$t^{\Pi} \equiv \frac{8af}{4bf + d(H+L)},$$

which is larger than t^*.

Therefore, when, for $\lambda \neq 1/2$, a skilled worker finds it individually rational to move to the larger region (i.e., when $(C_A' + w_A') - (C_B' + w_B') > 0$ as implied by $t < t^*$). In doing so, the worker neglects the collective welfare losses caused to the unskilled as expressed by $(C_A'' + C_B'')L/2$. He also neglects the negative effect on the welfare of the skilled expressed by

$$(C_A'' + C_B'')H/2 + (w_A'' + w_B'')H/2 \tag{9.70}$$

in which both terms are negative because $t^* < t^{\Pi}$. Hence, even within the skilled population the gainers cannot compensate the losers. This is because of the following general equilibrium effect: at the margin, fiercer price competition in the larger region depresses operating profits, thus wages, more than it increases them in the smaller region.

By giving weight only to the first positive term of (9.69), the skilled mover always imposes a negative external effect on the economy as a whole. In addition, because the absolute value of (9.70) decreases with t as transportation costs fall, the negative external effects become weak enough to align the agglomerated market equilibrium with the second best outcome. This suggests why a planner would choose agglomeration over a smaller interval of transportation costs than the market does.

9.5 ON THE IMPACT OF FORWARD-LOOKING BEHAVIOR

So far, we have assumed that workers care only about their current utility level, thus implying that only history matters. This is a fairly restrictive assumption to the extent that migration decisions are typically made on the grounds of current and future utility flows and various costs as a result of search, mismatch, and homesickness. In addition, this approach has been criticized because it is not a priori consistent with fully rational, forward-looking behavior. It is therefore important to determine if and how workers' expectations about the evolution of the economy may influence the process of agglomeration. In this section, we want to see how the core–periphery model can be used to shed more light on the interplay between history and expectations in the formation of the economic space when migrants maximize the intertemporal value of their utility flows. In particular, we are interested in identifying the conditions under which, when regions initially host different numbers of workers and firms, the common belief that workers will eventually agglomerate in the initially smaller region can reverse the historically inherited advantage of the larger region.

Somewhat different approaches have been proposed to tackle this problem, but they have yielded similar conclusions (Ottaviano 1999; Baldwin 2001; Ottaviano et al. 2002). In what follows, we use the model developed in Section 9.4.1 because it leads to a simple linear dynamic system that still allows for a detailed analysis of the main issues (Krugman 1991c; Fukao and Bénabou 1993). Formally, we want to determine the parameter domains of this model for which an equilibrium path exists consistent with workers' belief on the assumption that workers have perfect foresights (self-fulfilling prophecy).

Workers infinitely live with a rate of time preference equal to $\gamma > 0$. Because we wish to focus on the sole dynamics of migration, we assume that the consumption of the numéraire is positive for each point in time so that there is no intertemporal trade in the differentiated good.[28] For example, consider the case in which workers believe that region A will eventually attract the modern sector although region B is initially larger than A. Therefore, we want to test the consistency of the belief that, starting from time $\theta = 0$, all workers will end up being concentrated in A at some future date $\theta = T$, that is, there exists $T > 0$ such that, given $\lambda_0 < 1/2$,

$$\begin{aligned} &\dot{\lambda}(\theta) > 0 \qquad \theta \in [0, T) \\ &\lambda(\theta) = 1 \qquad \theta \geq T. \end{aligned} \tag{9.71}$$

Let $v_r(s)$ be the instantaneous indirect utility at time s in region r and denote by v^C the utility level incurred in region A when $\lambda = 1$. Given workers' expectations, we have

$$v_A(\theta) = v^C \quad \text{for all } \theta \geq T.$$

Because workers have perfect foresights, the easiest way to generate a non-bang-bang migration behavior is to assume that, when moving from one region to the other, workers incur a utility loss that depends on the rate of migration inasmuch as a migrant typically imposes a negative externality on the others (Mussa 1978). Specifically, we assume that the utility loss for a migrant at time θ is equal to $|\dot{\lambda}(\theta)|/\delta$, where $\delta > 0$ is a positive constant whose meaning is given below. Thus, under (9.71), the intertemporal utility at time 0 of a worker who moves from B to A at time $\theta \in [0, T)$ is assumed to be given by

$$\begin{aligned} U(\theta) &\equiv \int_0^\theta e^{-\gamma s} v_B(s)\, ds + \int_\theta^\infty e^{-\gamma s} v_A(s)\, ds - e^{-\gamma\theta}\dot{\lambda}(\theta)/\delta \\ &= \int_0^\theta e^{-\gamma s} v_B(s)\, ds + \int_\theta^T e^{-\gamma s} v_A(s)\, ds - e^{-\gamma T} v^C/\gamma - e^{-\gamma\theta}\dot{\lambda}(\theta)/\delta, \end{aligned}$$

where the first term stands for the utility accumulated in region B before moving to A, the second term for the utility obtained in A after migration, and the last represents the migration cost at time θ.

Because in equilibrium the skilled residing in region B do not want to delay their migration beyond T, it follows that[29]

$$\lim_{\theta \to T} |\dot{\lambda}(\theta)|/\delta = 0.$$

When θ tends toward T, we therefore obtain

$$U(T) = \int_0^T e^{-\gamma s} v_B(s)\, ds + e^{-\gamma T} v^C/\gamma.$$

Let

$$V_r(\theta) \equiv \int_\theta^T e^{-\gamma(s-\theta)} v_r(s)\, ds + e^{-\gamma(T-\theta)} v^C/\gamma \quad \theta \in [0, T) \tag{9.72}$$

be the discounted sum of future utility flows gross of moving costs of a worker *currently* (i.e., at time θ) residing in region $r = A, B$. Then, we have

$$U(\theta) - U(T) = \int_{\theta}^{T} e^{-\gamma s}[v_A(s) - v_B(s)]ds - e^{-\gamma\theta}\dot{\lambda}(\theta)/\delta$$
$$= e^{-\gamma\theta}[V_A(\theta) - V_B(\theta)] - e^{-\gamma\theta}\dot{\lambda}(\theta)/\delta$$
$$= e^{-\gamma\theta}\Delta V(\theta) - e^{-\gamma\theta}\dot{\lambda}(\theta)/\delta, \tag{9.73}$$

where $\Delta V(\theta) \equiv V_A(\theta) - V_B(\theta)$. Given that workers are free to choose when to migrate, it follows that

$$U(\theta) = U(T) \qquad \theta \in [0, T)$$

along the equilibrium path, and thus (9.73) implies that

$$\dot{\lambda}(\theta) = \delta\Delta V(\theta) \qquad \theta \in [0, T); \tag{9.74}$$

therefore, the private marginal cost of moving equals its private marginal benefit at any time $\theta < T$. In this expression, δ represents the speed of adjustment.

Using (9.72), we obtain the second law of motion by differentiating the utility gap $V_A(\theta) - V_B(\theta)$ with respect to θ

$$\dot{\Delta V}(\theta) = \gamma\Delta V(\theta) - \Delta v(\theta) \qquad \theta \in [0, T)$$
$$= \gamma\Delta V(\theta) - C_M t(t^* - t)(\lambda(\theta) - 1/2) \qquad \theta \in [0, T), \tag{9.75}$$

where $\Delta v(\theta) \equiv v_A(\theta) - v_B(\theta)$ is given by (9.60) for all $\theta \in [0, T)$, whereas $\Delta v(\theta) = 0$ for all $\theta \geq T$ because, as already noted, all workers expect a utility flow $v_A(\theta)$. As argued by Ottaviano (1999), expression (9.75) states that, during the migration process, the "annuity value" of being in A rather than in B (i.e., $\gamma\Delta V$) equals the "dividend" (Δv) plus the "capital gain" ($d\Delta V/d\theta$).

As a result, we obtain a system of two linear differential equations instead of one (see (9.60)), with the terminal conditions $\lambda(T) = 1$ and $\Delta V(T) = 0$. Because $\lambda = 1/2$ implies $\Delta V = 0$, the system (9.74) and (9.75) always has an interior steady state at $(\lambda, \Delta V) = (1/2, 0)$, which corresponds to the symmetric configuration. When $t > t^*$, it is the only steady state that is globally stable; hence, for the assumed belief (9.71) to be consistent with equilibrium, transportation costs must be low ($t < t^*$) – a case on which we concentrate from now on.

To identify the conditions under which expectations matter, we must study the global stability of the system (9.74) and (9.75). Because this system is linear, local and global stability properties coincide, thus allowing us to appeal only to the former. The eigenvalues of the Jacobian matrix system (9.74) and (9.75) are given by

$$\frac{\gamma \pm \sqrt{\gamma^2 - 4\delta C_M t(t^* - t)}}{2}. \tag{9.76}$$

When $t < t^*$ two scenarios may arise. In the first one, $\gamma > \sqrt{C_M \delta} t^*$, which is more likely to occur when moving costs are high, when consumers are impatient, or both. The two eigenvalues are still real but both positive. The steady state $(1/2, 0)$ is an unstable node, and there are two trajectories that steadily go to the endpoints, $(0, 0)$ or $(1, 0)$, depending on the initial spatial distribution of workers. In this case, only history matters: because $\lambda_0 < 1/2$ by assumption, there is a single trajectory that goes to $\lambda = 0$, as in the case in which the dynamics are given by (9.60). This implies that belief (9.71) is inconsistent with the equilibrium path.[30]

Matters turn out to be quite different in the second scenario in which $\gamma < \sqrt{C_M \delta} t^*$, that is, when both transportation and moving costs are low. Given that $C_M t(t^* - t) = 0$ at both $t = 0$ and $t = t^*$, the equation $C_M t(t^* - t) - \gamma^2/4\delta = 0$ has two positive real roots in t, denoted t_1^e and t_2^e, which are both smaller than t^*:

$$t_1^e \equiv \frac{t^* - E}{2} \quad \text{and} \quad t_2^e \equiv \frac{t^* + E}{2},$$

where

$$E \equiv \sqrt{(t^*)^2 - \gamma^2/C_M \delta}$$

stands for the size of the domain of values of t for which expectations matter; it shrinks as the discount rate γ increases or as the speed of adjustment δ decreases. Indeed, for $t \in (0, t_1^e)$ as well as for $t \in (t_2^e, t^*)$, both eigenvalues are real positive numbers, and the steady state $(1/2, 0)$ is an unstable node as before.

However, for $t \in (t_1^e, t_2^e)$, the eigenvalues become complex numbers with a positive real part so that the steady state is an unstable focus. Consequently, as illustrated in Figure 9.6, for any λ_0 close to but smaller than $1/2$, there is one trajectory going to $\lambda = 1$.[31] It is in such a case that expectations decide along which trajectory the system moves, and thus belief (9.71) is self-fulfilling. As shown in Figure 9.6, a symmetric trajectory going to $\lambda = 0$ exists under the symmetric belief. In other words, expectations matter for λ_0 close enough to $1/2$, whereas history matters otherwise.

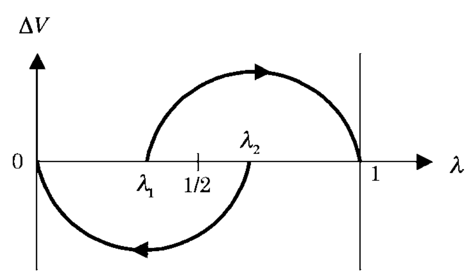


Figure 9.6: The overlap when expectations matter.

The range of values for which expectations matter, called the *overlap* by Krugman (1991c), can be obtained as follows. As observed by Fukao and Bénabou (1993), the system must be solved backwards in time starting from the terminal points (0, 0) and (1, 0). The first time the backward trajectories intersect the locus $\Delta V = 0$ allows the endpoints of the overlap to be identified:

$$\lambda_1 \equiv \frac{1}{2}(1 - \Lambda) \quad \text{and} \quad \lambda_2 \equiv \frac{1}{2}(1 + \Lambda)$$

where

$$\Lambda \equiv \exp\left(-\frac{\gamma\pi}{\sqrt{4\delta C_M t(t^* - t) - \gamma^2}}\right)$$

is the width of the overlap, which is an interval centered around $\lambda = 1/2$.

The overlap is nonempty as long as $t \in (t_1^e, t_2^e)$. Thus, the width of the overlap is increasing in δ, C_M, and t^*, whereas it decreases with γ. Moreover, it is ∩-shaped with respect to t, reaching a maximum at $t = t^*/2$. Consequently, we have shown the following result:

Proposition 9.6 *Let* λ_0 *be the initial spatial distribution of workers. If* $t < t^*$ *and* $\gamma < \sqrt{C_M \delta} t^*$*, then* $t_1^e \in (0, t^*/2)$*,* $t_2^e \in (t^*/2, t^*)$*,* $\lambda_1 \in (0, 1/2)$*, and* $\lambda_2 \in (1/2, 1)$ *exist, such that workers' beliefs about their future utility flows influence the process of agglomeration if and only if* $t \in (t_1^e, t_2^e)$ *and* $\lambda_0 \in [\lambda_1, \lambda_2]$.

Hence, history alone matters when t is large enough or small enough. In other words, the agglomeration process evolves as if moving costs were zero when obstacles to trade are high or low. By contrast, *when transportation costs take intermediate values and regions are not initially too different, the region that becomes the core is determined by workers' expectations and not by history.*

The range (t_1^e, t_2^e) may be explained as follows. Suppose, indeed, that the economy is such that $\lambda_0 < 1/2$ and ask what is needed to reverse an ongoing agglomeration process leading toward $\lambda = 0$. If the evolution of the economy were to change direction, workers would experience falling, instantaneous, indirect utility flows for some time period as long as $\lambda < 1/2$. The instantaneous, indirect utility flows would start growing only after λ became larger than $1/2$. Put differently, workers would first experience utility losses followed by utility gains. Because the losses would come before the gains, they would be less discounted. This reasoning provides the root for the intuition behind Proposition 9.6. When circular causation leads to substantial wage increases (that is, for intermediate values of t), the benefits of agglomerating at $\lambda = 1$ can compensate workers for the losses they incur during the transition phase, thus making the reversal of migration possible. On the contrary, for low or high values of t, the benefits of agglomerating at $\lambda = 1$ do not compensate workers for the losses.

Finally, when $\lambda(0)$ is closer to one endpoint, the period over which the skilled bear losses is larger, a higher rate of time preference giving more weight to them. Similarly, large moving costs (δ small) extend the time period over which workers' well-being is reduced, thus strengthening the weight of history in the agglomeration on process.

9.6 CONCLUDING REMARKS

Although the details of the agglomeration process vary with the model and the pricing policy, both the CES utility–iceberg transport cost and the quadratic utility–linear transport cost models of monopolistic competition show that a tendency toward agglomeration exists when transportation costs are sufficiently low. This finding suggests that the secular fall in transportation costs has intensified the geographical concentration of economic activities and confirms earlier analyses by Bairoch (1965) and Kaldor (1970, 340), who has observed the following:

> As communication between different regions becomes more intensified (with improvements in transport and marketing organisation), the region that is initially more developed industrially may gain from the progressive opening of trade at the expense of the less developed region whose development will be inhibited by it.

The decline in transport costs could have suggested that firms would become indifferent about their location. However, footloose activities, which are inherently independent of first nature, rely increasingly on second nature mechanisms. There are at least two reasons behind this phenomenon. First, as transportation costs decrease, firms have an incentive to concentrate their production in a smaller number of sites to better exploit scale economies in production. Second, low transportation costs tend to make price competition fierce, thus inducing firms to differentiate their products to relax price competition. This in turn leads firms to seek locations in which they have the best access to the largest pool of potential customers. In view of the results discussed in Chapters 7 and 8, this tendency suggests that *product differentiation is a powerful force toward agglomeration.*[32]

The process of economic unification within the European Union (EU) makes it ideal for testing the validity of the core–periphery model's predictions. As tariff barriers have been dropped and nontariff barriers reduced by a systematic policy aimed at improving integration of national markets, trade costs have dramatically decreased inside the EU. Recent empirical works seem to confirm the basic result of the core–periphery model. From 1968 to 1990, Amiti (1998) observed both an increase in the geographical concentration of economic activities within most EU state members and, for a vast majority of sectors, a tendency toward more agglomeration within the EU as a whole. Similarly, Brülhart (1998) noted an increase in the geographical concentration for 14 out of 18 industrial sectors between 1980 and 1990 within the EU. The cross-sectional analysis conducted by Haaland et al. (1999) in 13 European countries evidences cumulative causation in the sense that agglomeration of production and agglomeration of expenditure influence each other. Interestingly, these authors have also found that the most important determinant in the European geography is the location and size of demand, thus confirming a result obtained by Justman (1994) for the United States.

To our knowledge, one of the most elaborated existing studies has been conducted by Combes and Lafourcade (2001) who performed a structural estimation of a multiregional, multisectorial model with vertical linkages. More precisely, they consider 71 industrial sectors and 341 employment areas in France; transport costs are evaluated by means of a distance and time cost function in 1978 and 1993, built from the road network, gas price, highways tolls, and carriers' contracts. Their work shows that a decrease of about 40% in road transport costs was associated with a strengthening in regional specialization and inequalities.

Yet, these tentative conclusions must be qualified. Indeed, the introduction of various types of modifications in the core–periphery model casts doubt on the monotonicity of the agglomeration–transportation costs relationship. First, Venables (1996) noticed that the core–periphery structure rests very much on the assumption of spatial mobility of the skilled workers. This may be a reasonable assumption within some parts of the world (e.g., the United States) but not necessarily in others (e.g., the European Union). It is then more realistic to assume that all workers stay put. In this case, we have seen in Section 9.3 that the existence of an intermediate sector is another reason for a core–periphery structure. However, Krugman and Venables (1995) have invited us to reconsider the conclusions of Krugman (1991a) in that the relationship between the agglomeration rate of the modern sector and the level of transportation costs would be ∩-shaped instead of monotone decreasing. Because of the immobility of workers, the agglomeration of firms into a single region intensifies competition on the corresponding regional labor market. The tendency of firms to attract workers from the traditional sector in their region but not from the other region turns out to be a dominant centrifugal force when transportation costs are low enough. Thus, if agglomeration occurred for intermediate values of the transport costs, it would collapse eventually when these costs decreased further. More generally, Puga (1999) has shown that, if workers do not move between regions when their mobility costs between sectors are relatively low, a drastic fall in transportation costs eventually leads to geographic dispersion.

Second, when transport costs for the $\mathbb{M}$-goods (e.g., the industrial goods) are low relative to those for the $\mathbb{T}$-goods (e.g., the agricultural goods), the (relative) price of the $\mathbb{T}$-goods may increase faster than the price of the $\mathbb{M}$-goods, and thus the real income does not necessarily rise in the large region. This makes the concentration of activities into a single region less attractive for both workers and firms, which may lead to the collapse of the core–periphery structure (Calmette and Le Pottier 1995; Fujita et al. 1999, chap. 7). However, although transportation costs of both types of goods have declined since the beginning of the Industrial Revolution, it appears that what matters ultimately for the regional distribution of economic activities is not only the absolute levels of transport costs but also their relative values (Kilkenny 1998).

And, finally, Helpman (1998), Tabuchi (1998), and Ottaviano et al. (2002) have shown that positive urban costs, which take the form of housing and commuting costs (see Chapter 3), may also generate a ∩-shaped relationship. The reason is that a high concentration of workers into a single urban area gives rise to high land rents, as seen in Chapter 3, thus encouraging the relocation of workers and firms into the peripheral region. In these three cases, what fosters redispersion is the existence of factor price differentials in favor of the periphery that will trigger a new dispersion of activities once transport costs have sufficiently declined.

It is worth noting, however, that such models neglect possible firm relocations to the suburbs of the core metropolitan area rather than the periphery, as discussed in Sections 6.6 and 7.5. Indeed, suburbanization allows firms to enjoy low land rents and wages while preserving a high accessibility to a center supplying strategic inputs (e.g., a CBD). In fact, both urban and suburban locations seem to have been observed for a long time in many parts of the world. For example, in the case of Europe, Hohenberg, and Lees (1985, 129) observed the following:

> Without question, the long-run trend was for manufacturing to shift away from urban to rural locations, or at least to expand more in the latter, although cyclical changes in the economy could affect the trend. But in some cases, an urban location was always desirable, for example, where industries involved relatively heavy use of capital – fixed, circulating, or human – or the need for close entrepreneurial control. Less abstractly, the need to respond quickly to fashion or to other market changes, as well as the use of skilled labor, valuable raw materials, or expensive equipment argued for an urban location.

Moreover, all the models proposed in this chapter assume a given and fixed set of economic activities. However, as mentioned in Chapter 1, the decline of the industrial sector within big cities does not necessarily signify their economic and social decline. The continuous decrease in communications and transportation costs gives rise to new information-oriented economic activities that are typically developed within large metropolises (Feldman and Audretsch 1999; Henderson 1997b; Duranton and Puga 2002; Varga 2000).

In summary, for us, it is far from obvious that large metropolitan areas and regions are going to decline as transportation and communications costs keep decreasing.

APPENDIX

The following proof is due to Frédéric Robert-Nicoud. For notational convenience, set

$$\phi \equiv \Upsilon^{-(\sigma-1)} \tag{A.1}$$

and denote respectively by ϕ_{break} and ϕ_{sustain} the values of ϕ corresponding to

Υ_{break} and $\Upsilon_{\text{sustain}}$. Since $\sigma > 1$, ϕ is inversely related to Υ and belongs to the interval (0, 1].

Using (9.29) and (A.1), it is easy to show that ϕ_{sustain} is a solution to the equation

$$\phi^{-\mu\rho-1}[(1+\mu)\phi^2 + 1 - \mu] - 2 = 0.$$

Setting

$$f(\phi) \equiv \phi^{-\mu\rho-1}[(1+\mu)\phi^2 + 1 - \mu] - 2$$

it is straightforward that $f(0) \to \infty$ and $f(1) = 0$. Furthermore, it is readily verified that $f(\phi)$ has a single minimizer given by

$$\phi^* = \left[\frac{(1-\mu)(1+\mu\rho)}{(1+\mu)(1-\mu\rho)}\right]^{1/2} < 1.$$

Finally, it can be shown that $f'(0) < 0$ and $f'(1) > 0$. When ϕ varies from 0 to 1, all these facts put together imply that $f(\phi)$ decreases from arbitrarily large values to reach its minimum at $\phi^* > \phi_{\text{sustain}}$ since $f(\phi^*) < 0$ and, then, increases but takes negative values to reach the value 0 at $\phi = 1$.

Using (9.30) and (A.1), we obtain

$$\phi_{\text{break}} = \frac{(\sigma - 1 + \sigma\mu)(1+\mu)}{(\sigma - 1 - \sigma\mu)(1-\mu)}.$$

Evaluating $f(\phi)$ at ϕ_{break} reveals that $f(\phi_{\text{break}}) < 0$. Since $f(\phi_{\text{sustain}}) = 0$ and $f(\phi)$ decreases over $(0, \phi^*)$, it must be that

$$\phi_{\text{break}} > \phi_{\text{sustain}},$$

which yields the desired inequality

$$\Upsilon_{\text{break}} < \Upsilon_{\text{sustain}}.$$

NOTES

1. An earlier analysis that anticipated several aspects of Krugman's work was developed by Faini (1984). Ideas close to economic geography have already appeared in Krugman (1979) but were not fully worked out. Casetti (1980) has presented a simple dynamic analysis that permits the emergence of various equilibria and catastrophic transitions in a two-region economy. His article was written in 1970 but published in 1980, long before the emergence of new economic geography. Following a different approach, Arthur (1990; 1994, chap. 4) has shown the importance of "positive feedbacks" and history for the formation of regional clusters. Finally, in the presence of technological advantage of one region over the other, Markusen (1983) has shown that trade liberalization may generate more incentives

for labor migration, thus implying a partial agglomeration of workers within one region.

2. Yet, one should not forget that space allows the existence of pure profits even in free-entry equilibrium (Eaton and Lipsey 1978).
3. For more details, see Vives (1990).
4. This observation had already been made by Eaton and Lipsey (1977, 78–9).
5. This had already been noticed by Casetti (1980).
6. There is an older literature in trade theory that aims at modeling the transport sector just as another industry instead of using the iceberg approach. According to Neary (2001), this approach never led to simple results and is largely forgotten. However, this is enough to cast doubt on the generality of the results derived under the iceberg assumption because the two approaches do not seem to yield comparable results.
7. The most that can be obtained within this framework without resorting to numerical solutions probably has been achieved by Puga (1999).
8. When the number of firms is an integer, there are strategic interations (d'Aspremont, Dos Santos Ferreira, and Gérard-Varet 1996).
9. This agrees with the results obtained in industrial organization when firms are free to choose their product (Anderson et al. 1992, chap. 8).
10. It is worth noting that the summation in (9.11) may be interpreted as a market potential of region r. Indeed, each term in (9.11) gives the demand in s for a variety produced in r, which depends positively on the local income, negatively weighted by the accessibility of this region to region s. Finally, the competitiveness of region r is represented by $p_r(i)^{-\sigma}$.
11. We made a similar assumption in Section 8.4.
12. This process can be interpreted as a spatial Marshallian adjustment.
13. See Fujita et al. (1999, chap. 5) for more details.
14. Using the fact that function (9.27) always has a negative slope at $\Upsilon = 1$, we can show the shape of the curve as depicted in Figure 9.2.
15. It is worth noting that this result agrees with what we saw in Chapter 7 in the context of oligopolistic, but partial, equilibrium models.
16. This terminology is dictated by the secular fall in transportation costs.
17. We now assume that the modern sector uses labor through a Cobb–Douglas production function.
18. When $\beta = \delta$, (8.25) degenerates into a standard quadratic utility. See Ludema and Wooton (2000) for a model of economic geography with such a utility and quantity-setting firms.
19. This assumption allows one to focus on interior solutions only. Doing so has some costs in terms of generality but, as will be seen, larger benefits in terms of simpler analysis. The assumption is also consistent with the idea that each individual is interested in consuming both types of goods.
20. In the context of Section 9.2.5, this amounts to rescaling the demand intercept a in (9.52) and entails no loss of generality.
21. This is similar to what we obtained in Chapter 8 in the case of externalities. When the externality parameters are zero (ε in Section 9.2 and θ in Section 9.4), symmetry is the only equilibrium outcome. Sonnenschein (1982) has shown, to the contrary, a related result: perfect competition in each location is unable to explain

agglomerations. More precisely, if the initial distribution of firms is uneven along a given circle, then the spatial adjustment of firms in the direction of higher profit leads the economy toward a uniform long-run equilibrium in which each local economy is perfectly competitive. Mossay (2001) has revisited this problem in the context of local exchange economies and has shown that the spatial adjustment of consumers in the direction of higher utility also leads the economy toward a uniform long-run equilibrium when goods are gross substitutes. By contrast, when local markets are monopolistically competitive, Mossay (2001) has also shown that the spatial adjustment of workers renders the uniform long-run equilibrium unstable and, instead, generates local agglomerations.

22. It may also be shown that $\partial t^*/\partial\delta < 0$, which confirms what we just said.
23. Note that very small d implies that $t_{\text{trade}} < t^*$, and thus agglomeration always arises under trade.
24. Note that this comparison is made for a given level of transport costs. Starting from any τ for which equilibrium involves dispersion and lowering transport costs so that agglomeration arises in A, it can be shown that a sufficiently large decrease in τ also makes the unskilled in B better off.
25. Recall that we assume that transport costs are resource costs and not tariffs.
26. Although the argument above seems to rest very much on the linearity of demands, it is likely to hold true for any functional form such that the elasticity of demand for a variety increases with its price. This happens for any demand function that is concave or convex but less convex than a negative exponential – a condition often met in spatial pricing theory (Greenhut et al. 1987, 25). A notable violation of this property is given by isoelastic demands.
27. Under perfect competition and constant returns, those who remain in the source region are worse off when migration occurs. However, the global welfare always rises, thus allowing those left behind to be compensated by the migrants and the native-born in the destination region (Razin and Sadka 1997). As we have just seen, this is not necessarily true under increasing returns and monopolistic competition.
28. In the Krugman version of the core–periphery model, this is equivalent to assuming that expenditure equals income at each time (Baldwin 2001; Ottaviano 1999).
29. This condition has been introduced by Fukao and Bénabou (1993).
30. In this case, expectations do not play any role, and thus the myopic adjustment process studied in the previous section provides a good approximation of the qualitative evolution of the economy under forward-looking behavior (see also Baldwin 2001).
31. This trajectory is part of a spiral originating at $\lambda = 1/2$.
32. Similar arguments apply to labor: wage competition is a centrifugal force as is price competition, and a better access to a diversified labor pool for both firms and workers is a centripetal force.

PART IV

URBAN SYSTEMS AND REGIONAL GROWTH

10

Back to Thünen:

The Formation of Cities in a Spatial Economy

10.1 INTRODUCTION

In this book, we have studied cities from several different perspectives, for they constitute the most visible and important facet of the phenomenon of economic agglomeration. However, we have left untouched one important issue, namely, the location of cities. Under the assumption of costless intercity trade, cities are like floating islands, and their location is irrelevant. For many purposes, this has been a convenient simplifying assumption. However, because the central concern of this book is to bring back space into economics in its many aspects and dimensions, we cannot end our quest without exploring the question of where cities are established, and why. More important, such issues are likely to be crucial for the future of our economies. In an increasingly borderless world economy, the location of prosperous and growing cities should increasingly become a critical factor in the determination of people's well-being.

If the location of cities in the real world were arbitrary, it would be hopeless (and useless) to develop a theory about the location of cities. The reality, however, is quite the opposite. In fact, over the past century, economic geographers and historians have tirelessly advocated the surprising regularity in the actual structure of urban systems observed throughout the world, and so at different time periods (see, e.g., J. Marshall 1989, chap. 5; Hohenberg and Lees 1985, chap. 2). This effort has culminated in what is called "central place theory," as pioneered by Christaller ([1933] 1966) and Lösch ([1940] 1954). The aim of these authors was to explain the spatial distribution of economic activities within *a hierarchical system of urban centers.* Specifically, different goods, characterized by nested market areas and indexed accordingly by $i = 1, \dots, n$, are supplied by different firms. The goal is then to show that a location where good i is available also accommodates firms supplying all goods of order lower than i. The bulk of central place theory has been directed toward identifying conditions under which such a superposition of regular structures is possible. However, this theory has focused far too much on geometric considerations.[1] Indeed, there are no economic forces in these models that lead firms selling different goods to cluster, and thus it is hard to see why central places should emerge.

To the best of our knowledge, Eaton and Lipsey (1982) were the first to model *multipurpose shopping* as an economic foundation for the existence of clusters in which different goods are supplied by different firms.[2] Indeed, it is a well-documented fact that consumers organize their trips to satisfy various needs (see Thill and Thomas 1987 for a survey showing the empirical relevance of such a behavior). For example, on the same trip, a consumer buys different goods, meets friends, visits a movie theater, goes to the post office, or just wanders and looks around. That consumers group their purchases to reduce travel costs creates demand externalities that firms can exploit by locating with firms selling other goods. Or, as Stahl (1987, 790) put it,

> the market demand for a given commodity is not only dependent on the consumers' preferences, but also on the local supply of all other commodities and the conditions under which they are offered, such as their suppliers' advertising outlays, or their reputation in terms of selecting and marketing commodities. At any rate, the consumers' economies of scope in transactions outlays induce complementarities in the market demand for commodities that are substitutes in the individual consumer's eyes.

And, indeed, Eaton and Lipsey have identified a set of conditions under which the only equilibria involve clusters in which firms selling good 1 or good 2, each bought at a different frequency, are located together. Using the same framework, Quinzii and Thisse (1990) have proven that the socially optimal configuration of firms always involves the clustering of firms selling goods 1 and 2. These results therefore confirm the initial intuition. However, the spatial competition approach to the formation of central places taken in such articles becomes very quickly intractable. The reason is that trip-chaining implies a particular structure of substitution between outlets. This structure is such that it is hard for a consumer to determine her optimal spatial structure of purchases because this requires solving a particularly difficult combinatorial problem (Bacon 1971). It is accordingly easy to imagine how firms' demands become complex and intricate. Hence, a need exists for other approaches. Furthermore, although central place theory offers several fundamental insights about the organization of the economic geography (Mulligan 1984), it is fair to say that it has been largely descriptive.[3] In this chapter, we intend to develop a microeconomic approach to the formation of urban systems in which the location of cities matters.

To this end, we go back to the origin of geographical economics, that is, Thünen's *Isolated State*. As discussed in Chapter 3, Thünen (1828) started his work by making the following assumption:

> Imagine a very large town at the center of a fertile plain which is crossed by no navigable river or canal. Throughout the plain the soil is capable of cultivation and of the same fertility. Far from the town, the plain turns into an uncultivated wilderness which cuts off all communication between this state and the outside world.

There are no other towns on the plain. The central town must therefore supply the rural areas with all manufactured products, and in return will obtain all its provision from the surrounding countryside (p. 7 of the English translation).

Although countless variations of the Thünian model have appeared since then, the literature has left aside a fundamental issue: Why should all manufactured goods be produced in a single city?[4] To the best of our knowledge, Fujita and Krugman (1995) have been the first to offer a model in which city and agricultural land use are endogenously determined, thus making the analysis of Thünen complete. This work not only fills in a gap in the literature but, more important, it also suggests a microeconomic approach that allows one to study the emergence of an urban system in the spatial economy. Indeed, the identification of conditions under which a monocentric economy is sustainable as an equilibrium leads quite naturally to the fundamental question: Where and when do new cities emerge?

In Section 10.2, we present the results obtained by Fujita and Krugman for the monocentric economy. The underlying structure is closely related to that of the core–periphery model studied in Section 9.2. That is, agglomeration arises from love for variety on the consumer side (see also Section 7.2). However, here all workers are assumed to be identical and perfectly mobile between locations and sectors (typically, agriculture and industry). The centrifugal force now lies in the existence of a land market in the traditional (agricultural) sector, for producing the agricultural good requires both land and labor, thus leading to the spatial dispersion of demand for the manufactured goods because farmers are also consumers of such goods. In addition, the location space is continuous and one-dimensional (the same assumption is made throughout this chapter). Such a setting allows a synthesis of the Thünian model and the S–D–S model of monopolistic competition used in economic geography. We will see that a monocentric economy is a spatial equilibrium provided that the population size does not exceed some threshold value depending on the parameters of the economy.

In Section 10.3, love for variety on the consumer side is replaced by product variety in intermediate goods as in Section 9.3 (see also Section 4.4). Using the work of Fujita and Hamaguchi (2001), we will show that a monocentric economy in which both the industrial and intermediate sectors are established within a single city emerges as a spatial equilibrium provided that the cost of shipping intermediate commodities is high relative to the cost of transporting the consumption goods. Although they bear some resemblance, this setting and the one studied in Section 10.2 are not identical, and we will discuss the main analogies and differences. By contrast, once the intermediate inputs become cheaper to move relative to the output of the final sector, but not too much, the intermediate sector remains agglomerated, whereas the final output is locally produced and consumed when the population of workers does not exceed some level. In this case, the process of agglomeration with intermediate commodities becomes almost the same as the one studied in Section 10.2.

Hence, in Sections 10.2 and 10.3, the total population cannot be too large for a monocentric configuration to occur. This suggests that, as the population grows, new cities emerge when some critical population threshold is reached. When the population keeps rising, still more cities will appear, and so on. Using the framework of Section 10.2, we make this idea more precise in Section 10.4 by proposing an evolutionary approach to central place theory, which has been proposed by Fujita and Mori (1997). The main result is that *similar cities are created at (more or less) equal distances as the population increases continuously.* This provides a formal proof of one of the key ideas of central place theory:

> The normal pattern of human settlements. . . is one in which town growth is primarily a response to the needs of an agricultural population (J. Marshall, 1989, p. 15).

Our emphasis on population growth as a major reason for urbanization is justified by historical examples, such as the urbanization of Western Europe and the United States in the twelfth and nineteenth centuries, respectively, which were both accompanied by a substantial population increase. As expected, we will see that population growth not only fosters the development of incumbent cities but also provides incentives to found new cities.

In this chapter, unlike what we saw in Chapters 4 and 5, there is here no agent, such as a land developer or a local government, whose job it is to organize a new city by coordinating the actions of firms and workers. City formation is the outcome of a process involving agents who do not plan a priori to create a city but rather pursue their own interests. In other words, a city appears here as a complex system whose existence is the result of a self-organizing process. This modeling strategy, together with the idea that population growth is the engine of urbanization, is very much in the spirit of the evolutionary approach proposed by Hayek (1988, chaps. 3 and 8) to explain the working of an economy.[5]

Throughout, we consider only one group of differentiated goods (either for consumption or for use as intermediate inputs). Consequently, all cities produce the same type of goods and have similar sizes. To generate a hierarchical urban system à la Christaller, one needs to introduce different groups of differentiated goods with (1) different preference intensities, (2) different transport costs, and (3) different technologies. This extension will be discussed in our concluding section.

In this chapter, a city is considered to be a clustering of firms and workers that has no spatial extension. This means that workers do not bear any urban costs because the land market plays a role in the agricultural sector only. This is a severe restriction. We believe, however, that the models presented here may be viewed as a first attempt at synthesizing various aspects of a spatial economy that have been studied in previous chapters. Such a progressive synthesis is needed to put our understanding of the space-economy on solid grounds.

10.2 CITY FORMATION UNDER PREFERENCE FOR VARIETY

To study the formation of cities in a continuous space, we modify accordingly the model described in Section 9.2. More precisely, because all workers are now assumed to be perfectly mobile, we introduce a new immobile factor, *land*, which is used as an input in the traditional sector. It is also convenient to interpret this sector as being the agricultural one, whereas the modern sector is the industrial one.

10.2.1 The Framework

Consider an unbounded, one-dimensional location space X along which land has the same fertility and the same unit density. As in Section 9.2, there are two sectors, the industrial ($\mathbb{M}$) and the agricultural ($\mathbb{T}$) sectors. Unlike Section 9.2, all workers in the economy are assumed to be identical and free to choose their location and occupation. Each worker is endowed with one unit of labor that can be used indifferently in the $\mathbb{M}$-sector or in the $\mathbb{T}$-sector. There are L workers in the economy and a group of landowners who, for simplicity, are assumed to live on their land holdings. This implies that land rents are consumed where they are created. All consumers (workers and landowners) have the same utility function given by (9.1) and (9.2), so that the demand functions are still given by (9.3), (9.4), and (9.5).

The agricultural good is produced using both land and labor by means of fixed technological coefficients such that one unit of output requires one unit of land and $c^{\mathbb{T}}$ units of labor. In the modern sector, the technology of a firm is still given by (9.7) in which the marginal labor requirement c is normalized to 1 as before. Transport costs are now assumed positive for both the agricultural and industrial goods. If one unit of the $\mathbb{T}$-good is shipped from $r \in X$ to $s \in X$, only a fraction, given by $\exp(-\tau^{\mathbb{T}}|r - s|)$, arrives at its destination. For the $\mathbb{M}$-sector, the fraction arriving at s is denoted by $\exp(-\tau^{\mathbb{M}}|r - s|)$. Both $\tau^{\mathbb{T}}$ and $\tau^{\mathbb{M}}$ are positive and increase when the corresponding transport costs rise.

Finally, because our space is continuous, all variables are now described by continuous functions of locations. For example, the nominal wage rate at location r is $w(r)$ instead of w_r.

10.2.2 The Formation of a Monocentric Economy

Recalling Thünen, let us imagine a monocentric economy such as the one depicted in Figure 10.1. In such a setting, the production of all industrial goods occurs in a single city located, by convention, at $r = 0$, and the agricultural area surrounding the city expands from $-r_{\text{b}}$ to r_{b}, where r_{b} stands for the frontier distance to be determined. The city exports all the industrial goods to its agricultural hinterland and imports all the produce needed by its inhabitants. The question we want to solve is, Under which conditions is this configuration a spatial equilibrium? To answer this question, we proceed in two steps. First, we determine

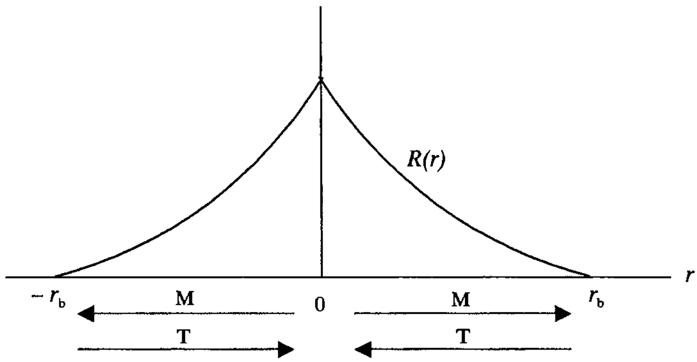


Figure 10.1: The monocentric economy.

the equilibrium prices of all goods, factors, and land rents, assuming that all industrial firms are located in the city. Then, we determine the conditions to be satisfied such that no industrial firm wants to locate away from the city. Because everything is symmetric, we restrict our analysis to the domain in which $r \geq 0$.

10.2.2.1 Equilibrium Prices

Let $p^{\mathbb{T}}(0)$ and $p^{\mathbb{M}}(0)$ be, respectively, the price of the agricultural good and the common price of the differentiated industrial goods at the city place. For the trade pattern between the city and its agricultural hinterland just described to arise, the price of each good at location $r \in X$ must be such that

$$p^{\mathbb{T}}(r) = p^{\mathbb{T}}(0)\exp(-\tau^{\mathbb{T}}r) \tag{10.1}$$

$$p^{\mathbb{M}}(r) = p^{\mathbb{M}}(0)\exp(\tau^{\mathbb{M}}r). \tag{10.2}$$

Let $R(r)$ and $w(r)$ be the land rent and wage rate prevailing at r. Inside the agricultural area, the land rent is equal to the price of the agricultural good minus the wage bill paid to the $c^{\mathbb{T}}$ workers needed to produce one unit of output:

$$R^*(r) = p^{\mathbb{T}}(r) - c^{\mathbb{T}}w(r). \tag{10.3}$$

At the agricultural border r_b, the land rent equals zero, and thus

$$w(r_b) = p^{\mathbb{T}}(0)\exp(-\tau^{\mathbb{T}}r_b)/c^{\mathbb{T}}. \tag{10.4}$$

Without loss of generality, we can set the wage in the city equal to 1:

$$w(0) = 1.$$

Turning to the industrial sector and restating (9.13) within the present context, we see immediately that, in the city, the equilibrium price of each variety is given by

$$p^{\mathbb{M}}(0) = \frac{1}{\rho}.$$

Let $L^{\mathbb{M}}$ be the mass of individuals working for the industrial sector in the city. Because an agricultural area of size $2r_b$ requires $2r_b c^{\mathbb{T}}$ workers, it follows that

$$L^{\mathbb{M}} = L - 2c^{\mathbb{T}} r_b.$$

Furthermore, given that the industrial goods are produced in the city only, applying (9.17), in which $\Upsilon_{rs} = \exp(\tau^{\mathbb{M}}|r - s|)$, we see that the price index of the industrial goods at location $r \in X$ is such that

$$\begin{aligned} P(r) &= p^{\mathbb{M}}(0)\left(\frac{L^{\mathbb{M}}}{\sigma f}\right)^{-1/(\sigma-1)} \exp(\tau^{\mathbb{M}} r) \\ &= \frac{1}{\rho}\left(\frac{L - 2r_b c^{\mathbb{T}}}{\sigma f}\right)^{-1/(\sigma-1)} \exp(\tau^{\mathbb{M}} r), \end{aligned} \tag{10.5}$$

which increases with distance from the city because of the increasing transport costs.

We are now ready to determine all the equilibrium prices. This can be achieved by using the following equilibrium conditions: (1) market clearing for the agricultural good and (2) the equality of the real wages at each location for all workers.

First, we notice that the income generated within the city equals $w(0)L^{\mathbb{M}} = L - 2r_b c^{\mathbb{T}}$. Hence, using (9.3) we obtain the demand for the agricultural good at the city place as follows:

$$D^{\mathbb{T}}(0) = (1 - \mu)\frac{L - 2r_b c^{\mathbb{T}}}{p^{\mathbb{T}}(0)}.$$

On the supply side, one unit of land at location r generates a gross return equal to $p^{\mathbb{T}}(r)$, thus giving rise to a demand for the agricultural good given by $(1 - \mu)p^{\mathbb{T}}(r)/p^{\mathbb{T}}(r) = (1 - \mu)$. Hence, μ units of the agricultural good are shipped to the city. Because only a fraction, $\exp(-\tau^{\mathbb{T}} r)$, of this amount reaches the city, the total supply of the agricultural good at the city place is

$$S^{\mathbb{T}}(0) = 2\mu \int_0^{r_b} \exp(-\tau^{\mathbb{T}} r)\, dr.$$

Thus, market clearing yields the following expression for the equilibrium price of the agricultural good at the city place:

$$p^{\mathbb{T}}(0) = \frac{1 - \mu}{2\mu} \frac{L - 2r_b c^{\mathbb{T}}}{\int_0^{r_b} \exp(-\tau^{\mathbb{T}} r)\, dr}. \tag{10.6}$$

Next, we move to the equality of real wages between farmers and workers. Equation (10.4) gives the nominal wage earned by a farmer residing at the agricultural frontier. Using (10.1), (10.4), and (10.5), the farmer's real wage is then as follows:

$$\omega(r_b) = w(r_b)[P(r_b)]^{-\mu}[p^{\mathbb{T}}(r_b)]^{-(1-\mu)}$$
$$= \frac{1}{c^{\mathbb{T}}}[P(0)]^{-\mu}[p^{\mathbb{T}}(0)]^{\mu}\exp[-\mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})r_b].$$

Because $w(0) = 1$, the real wage of a worker is

$$\omega(0) = [P(0)]^{-\mu}[p^{\mathbb{T}}(0)]^{-(1-\mu)}. \tag{10.7}$$

Hence, the equality of real wages between workers and farmers requires that

$$p^{\mathbb{T}}(0) = c^{\mathbb{T}}\exp[\mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})r_b]. \tag{10.8}$$

Putting (10.6) and (10.8) together shows that

$$L - 2c^{\mathbb{T}}r_b = \frac{2\mu c^{\mathbb{T}}}{1-\mu}\frac{1-\exp(-\tau^{\mathbb{T}}r_b)}{\tau^{\mathbb{T}}}\exp[\mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})r_b]. \tag{10.9}$$

Because the left-hand side of this expression decreases with r_b, whereas the right-hand side increases with r_b, it can readily be verified that (10.9) has a unique solution, which in turn allows for the determination of the other equilibrium values. It also follows from (10.9) that the equilibrium value of r_b increases from 0 to ∞ as the population L keeps rising from 0. This implies that we can treat L and r_b interchangeably in what follows.

Finally, introducing (10.5) and (10.8) into (10.7) and using (10.9), we obtain the equilibrium real wage common to farmers and workers:

$$\omega^* \equiv \omega(0) = k_1[1-\exp(-\tau^{\mathbb{T}}r_b)]^{\mu/(\sigma-1)}\exp[\mu(\rho^{-1}\mu-1)(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})r_b], \tag{10.10}$$

where

$$k_1 \equiv \rho^{\mu}(c^{\mathbb{T}})^{-(1-\mu)}\left(\frac{2\mu c^{\mathbb{T}}}{(1-\mu)\sigma f\tau^{\mathbb{T}}}\right)^{\mu/(\sigma-1)}.$$

Because $\omega^* = w(r)[P(r)]^{-\mu}[p^{\mathbb{T}}(r)]^{-(1-\mu)}$ must hold in equilibrium, the equilibrium nominal wage is as follows:

$$w^*(r) = \exp[\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}}]r. \tag{10.11}$$

10.2.2.2 Sustainability of the Monocentric Configuration

So far, we have assumed that all firms are located within the city. However, for the monocentric configuration to be a spatial equilibrium, we must show that no firm has an incentive to move away from the city and to set up in the countryside.

Let $w^{\mathbb{M}}(r)$ be the zero-profit wage rate that stands for the highest wage that a firm located at r is willing to pay given that the rest of the economy remains unchanged (this wage was formally introduced in Section 9.2 for a finite location space). The form of the equilibrium wage $w^*(r)$ given by (10.11) is such that no firm in the city has an incentive to move at r as long as $w^{\mathbb{M}}(r) \leq w^*(r)$. Instead of working directly with these two variables, it appears to be more convenient to use a monotonic transformation of their ratio.

Specifically, we define the *potential function of a firm* as

$$\Omega(r) = \left(\frac{w^{\mathbb{M}}(r)}{w^*(r)}\right)^{\sigma}. \tag{10.12}$$

The monocentric configuration is then a spatial equilibrium if and only if

$$\Omega(r) \leq 1 \quad \text{for all } r \geq 0$$

because there is no alternative location at which a zero-profit firm is able to offer more than workers actually earn. Since $w^*(r)$ is given by (10.11), it remains to determine $w^{\mathbb{M}}(r)$. Let $Y(0)$ be the total income generated by the manufacturing activities in the city and $Y(s)$ the total income generated by the farming activities at $r \leq r_b$. Then, using (9.19), we obtain

$$w^{\mathbb{M}}(r) = \kappa_2 \left\{ Y(0) \exp[-(\sigma - 1)\tau^{\mathbb{M}} r][P(0)]^{\sigma - 1} + \int_{-r_b}^{r_b} Y(s) \exp[-(\sigma - 1)\tau^{\mathbb{M}}|r - s|][P(s)]^{\sigma - 1} ds \right\}^{1/\sigma}, \tag{10.13}$$

where $\kappa_2 \equiv \rho[\mu/(\sigma - 1)f]^{1/\sigma}$

Clearly, we have $Y(0) = w^*(0)L^{\mathbb{M}} = L - 2r_b c^{\mathbb{T}}$ and $Y(r) = p^{\mathbb{T}}(r) = p^{\mathbb{T}}(0) \exp(-\tau^{\mathbb{T}} r)$, in which $p^{\mathbb{T}}(0)$ is given by (10.8). Thus, using (10.5) for the price index $P(r)$ as well as (10.9) for $Y(0)$, we may rewrite (10.13) as follows:

$$w^{\mathbb{M}}(r) = \left(\frac{\sigma f}{\rho^{\sigma - 1}}\right)^{1/\sigma} \kappa_2 \left\{ \exp[-(\sigma - 1)\tau^{\mathbb{M}} r] + \frac{1 - \mu}{2\mu} \frac{\tau^{\mathbb{T}}}{1 - \exp(-\tau^{\mathbb{T}} r_b)} \times \int_{-r_b}^{r_b} \exp[-\tau^{\mathbb{T}}|s| - (\sigma - 1)\tau^{\mathbb{M}}(|r - s| - |s|)]\, ds \right\}^{1/\sigma}. \tag{10.14}$$

Substituting (10.11) and (10.14) into (10.12) yields

$$\Omega(r) = \mu \exp\{-\sigma[\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}}]r\} \cdot \Bigg\{ \exp[-(\sigma-1)\tau^{\mathbb{M}} r] + \frac{1-\mu}{2\mu} \cdot \frac{\tau^{\mathbb{T}}}{1-\exp(-\tau^{\mathbb{T}} r_{\mathrm{b}})} \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} \exp[-\tau^{\mathbb{T}}|s| - (\sigma-1)\tau^{\mathbb{M}}(|r-s| - |s|)]\, ds \Bigg\}^{1/\sigma}. \tag{10.15}$$

As shown in the appendix, this expression can be rewritten in a more convenient way as follows:

$$\Omega(r) = \exp(-\eta r)\Bigg\{1 + (1-\mu)(\sigma-1)\tau^{\mathbb{M}} \cdot \int_0^r (\exp[2(\sigma-1)\tau^{\mathbb{M}} s]) \times \left(1 - \frac{1-\exp(-\tau^{\mathbb{T}} s)}{1-\exp(-\tau^{\mathbb{T}} r_{\mathrm{b}})}\right) ds\Bigg\} \tag{10.16}$$

in which r_{b} appears only once, whereas

$$\eta \equiv \sigma[(\mu+\rho)\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}}]. \tag{10.17}$$

Expression (10.16) is depicted in Figure 10.2 for the following set of values: $\rho = 0.75$ (hence, $\sigma = 4$), $\mu = 0.5$, $\tau^{\mathbb{T}} = 0.8$, $\tau^{\mathbb{M}} = 1$, and $c^{\mathbb{T}} = 0.5$. For any given value of r_{b}, (10.16) is represented by a curve called the *potential curve*. Because r_{b} is uniquely determined by the population size L, this amounts to saying that each potential curve is associated with a single value of L.

For the monocentric configuration to be an equilibrium, the potential function associated with the corresponding value of L should never exceed 1. To determine when this is so, we must study the behavior of (10.16). First, it is apparent that $\Omega(0) = 1$, which means that firms located within the city pay a wage equal to the zero-profit wage at $r = 0$. Thus, for the function $\Omega(r)$ not to exceed 1 in a neighborhood of $r = 0$, the slope of $\Omega(r)$ at $r = 0$ given by

$$\Omega'(0) = \sigma[(1-\mu)\tau^{\mathbb{T}} - \mu(1+\rho)\tau^{\mathbb{M}}] \tag{10.18}$$

must be nonpositive, that is,

$$\frac{1-\mu}{\mu(1+\rho)} \le \frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}}. \tag{10.19}$$

Observe that (10.18) is independent of L and is, therefore, the same for all potential curves.

If (10.19) holds as a strict inequality, the potential function has a cusp at $r = 0$ (see Figure 10.2), which corresponds to the *lock-in effect* generated by

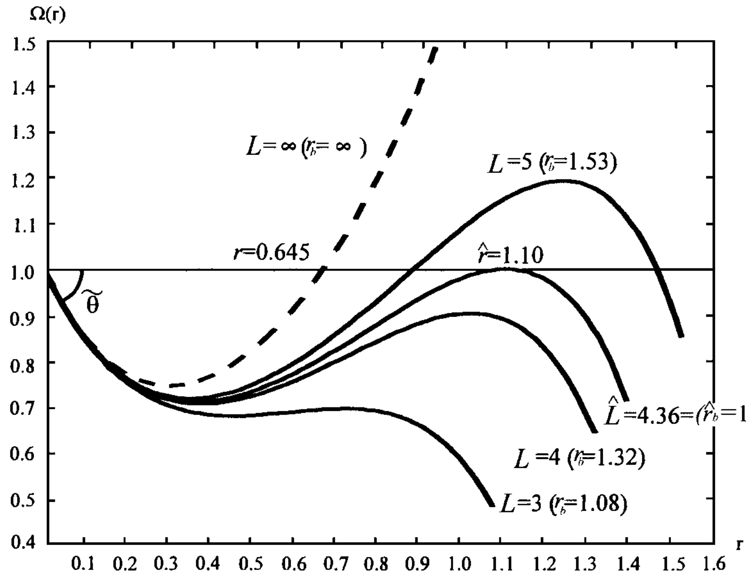


Figure 10.2: Potential curves for the monocentric configuration.

the agglomeration of firms and workers in the city. In this case, no firm finds it profitable to move a short distance away from the city because a sufficiently large share of its demand comes from there. As shown by (10.19), this happens when (1) the transport costs of the industrial goods are high relative to those of the agricultural good, (2) varieties are good substitutes (ρ is large) so that the elasticity of substitution is high, and (3) the expenditure share μ on the industrial goods is large.

However, (10.19) is a local sufficient condition for the monocentric configuration to be an equilibrium, not a global one. For that, we must return to (10.16). It is readily verified that the potential curve shifts upward (except at the origin) as r_b increases. Because r_b increases with L, this implies that an increase in the labor force shifts the corresponding potential curve upward. The explanation for this is that a larger population requires a larger agricultural area, and thus, when the other industrial firms are in the city, a firm moving to a location in the countryside expects a larger demand for its output. As a result, the potential function may well exceed 1 when L becomes sufficiently large.

To investigate such a possibility, we introduce the limit potential curve $\bar{\Omega}(r)$ associated with $r_b \to \infty$ (and, hence, $L \to \infty$) representing the upper limit of all potential curves. Taking the limit of (10.16) for $r_b \to \infty$ and rearranging terms, we obtain

$$\bar{\Omega}(r) = (1 - K)\exp(-\eta r) + K \exp \gamma r, \tag{10.20}$$

where K and γ are two constants defined as

$$K \equiv \frac{(1-\mu)\rho\tau^{\mathbb{M}}}{(1-\mu)\rho\tau^{\mathbb{M}} + (\rho-\mu)(\tau^{\mathbb{M}}+\tau^{\mathbb{T}}) - \Omega'(0)/\sigma}$$

$$\gamma \equiv \sigma(\rho-\mu)(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})$$

in which we have used (10.17) rewritten as follows:

$$\eta \equiv (1-\mu)(\sigma-1)\tau^{\mathbb{M}} - \Omega'(0).$$

Because $\bar{\Omega}'(0) = \Omega'(0)$ and we consider only the case in which $\Omega'(0) \leq 0$, the limit potential curve slopes down at the origin. Furthermore, since $\bar{\Omega}(r)$ is the sum of two exponential functions, it has at most one turning point (at which $\bar{\Omega}'(r) = 0$). Accordingly, $\bar{\Omega}(r)$ exceeds 1 for some $r > 0$ if and only if $\bar{\Omega}(\infty) > 1$. Because $\Omega'(0) \leq 1$ implies $\eta > 0$, it follows from (10.20) that $\bar{\Omega}(\infty) = K \exp(\gamma\infty)$, and thus $\bar{\Omega}(r)$ exceeds 1 for some positive r if and only if $K \exp(\gamma\infty) > 1$.

If $\rho < \mu$, then $\gamma < 0$ and, hence, $K \exp(\gamma\infty) = 0$. If $\rho = \mu$, then $\gamma = 0$ and $K \leq 1$, so that $K \exp(\gamma\infty) = K \leq 1$. Hence, when

$$\rho \leq \mu, \tag{10.21}$$

it follows that $\bar{\Omega}(r) \leq 1$ for all $r > 0$. This in turn implies that $\Omega(r) < 1$ for all $r > 0$. Condition (10.21) holds when either the industrial goods are very differentiated (ρ is small) or a large fraction of the labor force works in the city. In either case, the city is the most profitable site for the industrial sector however large the population L of the economy is. Note that (10.21) corresponds to the black hole condition discussed in Section 9.2.3: the city attracts all industrial activities and corresponds to *the* megalopolis.

Conversely, and more interestingly, if

$$\rho > \mu \tag{10.22}$$

then $\gamma > 0$, whereas $0 < K$ so that $K \exp(\gamma\infty) = \infty$, thus implying that $\bar{\Omega}(r)$ exceeds 1 for some value $r > 0$. In this case, as L keeps rising, the potential function $\Omega(r)$ will exceed 1 for some $r > 0$. Therefore, when varieties are not very differentiated, the city population is relatively small, or both, the agricultural area generates a sufficiently high demand for the industrial goods once L is large enough. This in turn implies that a firm finds it profitable to locate away form the city to benefit from some monopoly power in its local market.

Figure 10.2 illustrates the case in which the potential function is just equal to 1 at some critical distance $\hat{r} = 1.10$ when $L = \hat{L}(= 4.36)$. Any further increase in the population size above $\hat{L}$ makes the location $\hat{r}$ more profitable than the city for any single firm, thus suggesting that a new city might well emerge there. However, we postpone the analysis of this transition to Section 10.4. It is also worth noting the existence of an area of *urban shadow* in which the potential function never exceeds 1 (from $r = 0$ to $r = 0.645$), which means that a new city will never appear within this area.

Putting all those results together, we may conclude as follows:[6]

Proposition 10.1 *Consider a monocentric economy. For this configuration to be a spatial equilibrium, the following condition must hold:*

$$\frac{1-\mu}{\mu(1+\rho)} \leq \frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}}$$

1. *if the foregoing condition holds and if $\rho \leq \mu$, then the monocentric configuration is an equilibrium regardless of the population size; and*
2. *if the foregoing condition holds and if $\rho > \mu$, then a critical population level $\hat{L}$ exists such that the monocentric configuration is an equilibrium for all $L \leq \hat{L}$, whereas it is no longer an equilibrium as soon as $L > \hat{L}$.*

The present setting allows for some interesting comparative statics results studied by Fujita and Krugman (1995). Assume that technological progress in agriculture has led to the development of a labor-saving technology, such as those observed in the nineteenth century in the United Kingdom and later on, in the twentieth century, in the United States and then in Europe. This means that $c^{\mathbb{T}}$ decreases. Then, although the agricultural area expands, such a change in the agricultural technology fosters migration from the agricultural hinterland toward the city. This in turn leads to an increase in the urban population as well as to an expansion of the industrial sector through a wider product range and a larger mass of firms. Such a pattern not only concurs with historical evidence but also agrees with what we may observe today in countries experiencing industrialization.

Consider now a decrease in transport cost for either type of good (either $\tau^{\mathbb{M}}$ or $\tau^{\mathbb{T}}$ decreases). This renders distant locations more attractive, thus inducing migration from the urban center of the economy toward the agricultural frontier that moves further away. To some extent, this might provide a rationale for the westward expansion of the United States economy in the second half of the nineteenth century, during which transport costs fell dramatically as railways and waterways were built and developed.

10.2.3 Welfare in the Monocentric Economy

We now come to the impact of the population growth on the economy's welfare. In this perspective, one must keep in mind that the economy involves two different groups of agents, that is, workers and landlords. Workers' welfare is represented by their equilibrium real wage, as given by (10.10). Differentiating this expression with respect to r_b, we obtain

$$\frac{d\omega^*}{dr_b} = \frac{\mu\omega^*}{\sigma - 1}\left[\sigma(\mu - \rho)(\tau^{\mathbb{M}} + \tau^{\mathbb{T}}) + \frac{\tau^{\mathbb{T}}}{\exp(\tau^{\mathbb{T}} r_b) - 1}\right]. \tag{10.23}$$

Regarding the landlords' welfare, we know that they receive per unit of land a rent equal to $R(r) = p^{\mathbb{T}}(r) - c^{\mathbb{T}} w(r)$ for all $r \leq r_b$. As a result, their real income $\omega^L(r)$ is given by

$$\omega^L(r) = [p^{\mathbb{T}}(r) - c^{\mathbb{T}} w^*(r)][P(r)]^{-\mu}[p^{\mathbb{T}}(r)]^{-(1-\mu)} \quad \text{for all } 0 \leq r \leq r_b.$$

Because

$$\omega^* = w^*(r)[P(r)]^{-\mu}[p^{\mathbb{T}}(r)]^{-(1-\mu)}$$

it follows that

$$\omega^L(r) = \omega^*\left(\frac{p^{\mathbb{T}}(r)}{w^*(r)} - c^{\mathbb{T}}\right) \quad 0 \leq r \leq r_b. \tag{10.24}$$

Substituting (10.1) and (10.11) into (10.24) and using (10.8), we obtain

$$\omega^L(r) = \omega^* c^{\mathbb{T}}\{\exp[\mu(\tau^{\mathbb{M}} + \tau^{\mathbb{T}})(r_b - r)] - 1\} \quad 0 \leq r \leq r_b. \tag{10.25}$$

If the black hole condition holds ($\rho \leq \mu$), it is obvious from (10.23) that workers are always better off when the population rises. Furthermore, on the right-hand side of (10.25), the second factor increases with r_b and, therefore, with L. Consequently, the welfare of both workers and landlords always increases with the population size. This is because the population growth leads to a wider array of varieties that dominates the higher trade costs associated with a larger agricultural area (recall that varieties are very differentiated).

When the black hole condition does not hold ($\rho > \mu$), then the first term on the right-hand side of (10.23) is negative, whereas the second term decreases continuously with r_b (hence, with L) from infinity to zero. Accordingly, ω^* is $\cap$-shaped, thus implying that the real wage of workers first increases and then decreases with the population size. The highest welfare level is achieved for a particular population size L^o associated with the fringe distance

$$r_b^o = \frac{1}{\tau^{\mathbb{T}}}\left[\log\left(1 + \frac{1}{\sigma(\rho - \mu)}\frac{\tau^{\mathbb{T}}}{\tau^{\mathbb{M}} + \tau^{\mathbb{T}}}\right)\right], \tag{10.26}$$

which is obtained by setting the right-hand side of (10.23) equal to zero. This is so because the varieties are not sufficiently differentiated to prevent the increase in trade costs caused by the enlargement of the agricultural hinterland from becoming predominant. In this case, L^o is the worker-optimal population size. Because (10.26) decreases with ρ, the worker-optimal population size increases with product differentiation.

When L rises above L^o, the welfare of landlords may continue to increase, for their income, which comes from land rent, keeps rising as the agricultural area expands. More precisely, using (10.23) and (10.24), we find that

$$\begin{aligned}
\frac{1}{\omega^L(r)}\frac{d\omega^L(r)}{dr_\mathrm{b}} &= \frac{\mu}{\sigma-1}\left[\sigma(\mu-\rho)(\tau^{\mathbb{M}}+\tau^{\mathbb{T}}) + \frac{\tau^{\mathbb{T}}}{\exp(\tau^{\mathbb{T}}r_\mathrm{b})-1}\right] \\
&\quad + \frac{\mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})}{1-\exp[-\mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}})(r_\mathrm{b}-r)]} \\
&> \frac{\sigma\mu(\mu-\rho)}{\sigma-1}(\tau^{\mathbb{M}}+\tau^{\mathbb{T}}) + \mu(\tau^{\mathbb{M}}+\tau^{\mathbb{T}}) \\
&= \frac{\sigma\mu^2}{\sigma-1}(\tau^{\mathbb{M}}+\tau^{\mathbb{T}}) \\
&> 0, \qquad (10.27)
\end{aligned}$$

which holds for all r_b and, therefore, for all L.

We may then summarize the foregoing results as follows:

Proposition 10.2 *Consider a monocentric economy. Then, if the population size increases*

1. *the welfare of the landlords always increases within the agricultural area;*
2. *when $\rho \leq \mu$, workers' welfare always increases;*
3. *when $\rho > \mu$, workers' welfare increases to some population level and declines beyond.*

Hence, workers' welfare displays a pattern similar, although not identical, to that of the market equilibrium. Of course, the two patterns are not identical, but they have the same "shape." When the city is not a black hole ($\rho > \mu$), increasing the population beyond L^o entails a conflict between workers' and landlords' interests because rising L yields a decrease in workers' individual welfare. Since people are free to move, some workers will eventually be induced to move away from the city to create new cities together with some farmers, which is an issue that will be fully analyzed in Section 10.4.

10.3 CITY FORMATION WITH INTERMEDIATE COMMODITIES

In this section, we focus on the role of *variety in intermediate goods* in city formation by combining the ideas of Sections 9.3 and 10.2. To this end, we modify the model of Section 9.3 as we did with the model of 9.2 in the previous section. Thus, the resulting model is essentially the counterpart to that presented in Section 10.2. Land is an immobile input for the agricultural sector. The industry is formed by two sectors vertically linked, that is, the $\mathbb{M}$-sector producing the homogeneous $\mathbb{M}$-good for consumption and the $\mathbb{I}$-sector supplying a large variety of intermediate goods to the $\mathbb{M}$-sector. Such a setting yields a richer set of outcomes than before. In particular, we will see that two types of monocentric configurations exist involving very different patterns of trade. In the first, both sectors are agglomerated together and the resulting city exports the $\mathbb{M}$-good to the agricultural hinterland. In the second, the $\mathbb{I}$-sector is concentrated within a city; as for the $\mathbb{M}$-sector, it is partially concentrated in the city while the rest is mixed with the agricultural sector such that each place produces the $\mathbb{M}$-good for its own needs. The city now exports the intermediate goods only – a pattern that resembles that of several cities in developed economies. The first pattern is called an *integrated city*, and the second one an $\mathbb{I}$-*specialized city.* Not surprisingly, the former (respectively, the latter) tends to arise when the transportation costs of the intermediate goods are high (respectively, low). We will also show that, when the economy involves the agglomeration of both industrial sectors, the growth of the population alone can never destroy the equilibrium. By contrast, in the case of an $\mathbb{I}$-specialized city, population growth eventually leads to the formation of new cities.

After having described our new framework, we will proceed by studying the integrated and $\mathbb{I}$-specialized cities in turn.

10.3.1 The framework

The approach taken here is fairly similar to that followed in the previous section. In particular, space is given by a unbounded, linear space X, whereas land has the same fertility and density across locations. The agricultural (or $\mathbb{T}$-) sector produces one unit of the agricultural good using $c^{\mathbb{T}}$ units of labor and one unit of land. However, the $\mathbb{M}$-sector now produces a homogeneous $\mathbb{M}$-good under constant returns using labor and a continuum of intermediate inputs supplied by the $\mathbb{I}$-sector. The production function of the $\mathbb{M}$-sector is given by (9.34), yielding the unit production cost (9.40) as well as the input demands (9.41) and (9.42). As in Section 9.3, each intermediate variety is produced by using labor only according to the technology (9.33), which thus exhibits scale economies.

The variable Q in the utility (9.1) now stands for the homogeneous output of the $\mathbb{M}$-sector; the demands for this good and for the agricultural good are given by (9.37) and (9.38). As in the Section 10.2, we assume that transport costs have the iceberg form: if one unit of the agricultural good (respectively, the $\mathbb{M}$-good or the $\mathbb{I}$-good) is shipped from $r \in X$ to $s \in X$, only a fraction, given by $\exp(-\tau^{\mathbb{T}}|r-s|)$ (respectively, $\exp(-\tau^{\mathbb{M}}|r-s|)$ or $\exp(-\tau^{\mathbb{I}}|r-s|)$), arrives at destination, where $\tau^{\mathbb{T}}$ (respectively, $\tau^{\mathbb{M}}$ or $\tau^{\mathbb{I}}$) is a positive constant.

Finally, as in the Section 10.2, all workers are identical and free to choose their job in any of the three sectors as well as their location, either in the city or in the countryside.

The two equilibria we want to analyze share several similar features that are now described. Because scale economies arise only in the $\mathbb{I}$-sector, we restrict ourselves to the case in which the entire intermediate sector is established within the city. The agricultural area is assumed to surround the city symmetrically from $-r_{\mathrm{b}}$ to r_{b}. In both equilibria, let $p^{\mathbb{T}}(r)$ and $p^{\mathbb{M}}(r)$ be, respectively, the price of the agricultural good and the price of the $\mathbb{M}$-good at distance r from the city, whereas ω^* is the common equilibrium real wage earned by the workers in the two industrial sectors as well as by the farmers. Then, the equilibrium nominal wage at r is as follows:

$$w^*(r) = \omega^*[p^{\mathbb{M}}(r)]^{\mu}[p^{\mathbb{T}}(r)]^{(1-\mu)} \quad 0 \leq r \leq r_{\mathrm{b}}. \tag{10.28}$$

Using the normalization $w^*(0) = 1$, from (10.28) we obtain

$$\omega^* = [p^{\mathbb{M}}(0)]^{-\mu}[p^{\mathbb{T}}(0)]^{-(1-\mu)} \tag{10.29}$$

$$w^*(r) = [p^{\mathbb{M}}(r)/p^{\mathbb{M}}(0)]^{\mu}[p^{\mathbb{T}}(r)/p^{\mathbb{T}}(0)]^{(1-\mu)}. \tag{10.30}$$

Through a now standard argument with a normalization such that $w^*(0) = 1$, it is easy to show that the common equilibrium price of the intermediate commodities produced in the city is $p^{\mathbb{I}}(0) = w^*(0)/\rho = 1/\rho$. Hence, their delivered price at r is

$$p^{\mathbb{I}}(r) = \frac{\exp(\tau^{\mathbb{I}} r)}{\rho}. \tag{10.31}$$

Let $L^{\mathbb{I}}$ be the mass of workers in the $\mathbb{I}$-sector. Then, because the labor requirement of each firm under the zero-profit condition equals σf, $L^{\mathbb{I}}/\sigma f$ varieties are produced in the city. Using (9.43) and (10.31), the price index of the $\mathbb{I}$-varieties at location r can be obtained as in (10.5) and is given by the following expression:

$$P^{\mathbb{I}}(r) = \frac{1}{\rho}\left(\frac{L^{\mathbb{I}}}{\sigma f}\right)^{-1/(\sigma-1)} \exp(\tau^{\mathbb{I}} r). \tag{10.32}$$

From (9.40), the unit production cost of the $\mathbb{M}$-good at location r is now given by

$$\begin{aligned} c^{\mathbb{M}}(r) &= \alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}[w^*(r)]^{1-\alpha}[P^{\mathbb{I}}(r)]^{\alpha} \\ &= \kappa_3 (L^{\mathbb{I}})^{-\alpha/(\sigma-1)}[w^*(r)]^{1-\alpha}\exp(\alpha\tau^{\mathbb{I}} r), \end{aligned} \tag{10.33}$$

where

$$\kappa_3 \equiv \alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}\rho^{-\alpha}(\sigma f)^{\alpha/(\sigma-1)}.$$

Let $L_o^{\mathbb{M}}$ be the labor force working for the $\mathbb{M}$-sector in the city and $L^{\mathbb{M}}(r)$ the density of workers in the $\mathbb{M}$-sector at $r \neq 0$. It is then clear that the total city income equals $L_o^{\mathbb{M}} + L^{\mathbb{I}}$ because $w^*(0) = 1$, whereas the total income per unit of land at $r \neq 0$ is $p^{\mathbb{T}}(r) + w^*(r)L^{\mathbb{M}}(r)$. As a consequence, from (9.38), the demands for the $\mathbb{M}$-good in the city and at $r \neq 0$, respectively, given by

$$Q_o^{\mathbb{M}} = \mu\big(L_o^{\mathbb{M}} + L^{\mathbb{I}}\big)/p^{\mathbb{M}}(0) \tag{10.34}$$

$$Q^{\mathbb{M}}(r) = \mu[p^{\mathbb{T}}(r) + w^*(r)L^{\mathbb{M}}(r)]/p^{\mathbb{M}}(r). \tag{10.35}$$

On the supply side, from (9.41), the outputs of the $\mathbb{M}$-good produced in the city and at $r \neq 0$ are as follows:

$$X_o^{\mathbb{M}} = L_o^{\mathbb{M}}/(1-\alpha)c^{\mathbb{M}}(0) \tag{10.36}$$

$$X^{\mathbb{M}}(r) = w^*(r)L^{\mathbb{M}}(r)/(1-\alpha)c^{\mathbb{M}}(r). \tag{10.37}$$

Zero profits in the $\mathbb{M}$-sector imply the following two conditions:

$$L_o^{\mathbb{M}} > 0 \Rightarrow p^{\mathbb{M}}(0) = c^{\mathbb{M}}(0)$$

$$L^{\mathbb{M}}(r) > 0 \Rightarrow p^{\mathbb{M}}(r) = c^{\mathbb{M}}(r).$$

The market-clearing condition for the $\mathbb{M}$-good depends on the equilibrium type, and its study is postponed to the next two sections.

We now turn to the market-clearing condition for labor. We have

$$2c^{\mathbb{T}}r_{\mathrm{b}} + L_o^{\mathbb{M}} + \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} L^{\mathbb{M}}(r)\,dr + L^{\mathbb{I}} = L. \tag{10.38}$$

It turns out to be convenient to rewrite this expression. For that, note that the zero-profit condition in the $\mathbb{I}$-sector implies that its total cost, $L^{\mathbb{I}}$, is equal to its total revenue, which in turn is given by the total expenditure of the $\mathbb{M}$-sector on the $\mathbb{I}$-good as

$$\alpha\left[c^{\mathbb{M}}(0)X_o^{\mathbb{M}} + \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} c^{\mathbb{M}}(r)X^{\mathbb{M}}(r)\,dr\right],$$

which, from (10.36) and (10.37), is equal to

$$\alpha\left[\frac{L_o^{\mathbb{M}}}{1-\alpha} + \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} \frac{w^*(r)L^{\mathbb{M}}(r)}{1-\alpha}\,dr\right].$$

Therefore, it follows that

$$L^{\mathbb{I}} = \frac{\alpha}{1-\alpha}\left[L_o^{\mathbb{M}} + \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} w^*(r)L^{\mathbb{M}}(r)\,dr\right]. \tag{10.39}$$

Substituting (10.39) into (10.38), we may then rewrite the labor-market-clearing condition as follows:

$$2c^{\mathbb{T}}r_{\mathrm{b}} + \frac{L_o^{\mathbb{M}}}{1-\alpha} + \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} \left[1 + \frac{\alpha}{1-\alpha}w^*(r)\right]L^{\mathbb{M}}(r)\,dr = L. \tag{10.40}$$

Regarding the equilibrium location conditions for the $\mathbb{M}$-sector that is perfectly competitive, we have

$$p^{\mathbb{M}}(r) \leq c^{\mathbb{M}}(r) \quad r \geq 0. \tag{10.41}$$

To obtain these conditions for the $\mathbb{I}$-firms, as in (10.12) we define the potential function for an $\mathbb{I}$-firm as

$$\Omega(r) = \left(\frac{w^{\mathbb{I}}(r)}{w^*(r)}\right)^{\sigma} \tag{10.42}$$

in which $w^{\mathbb{I}}(r)$ stands for the zero-profit wage rate that an $\mathbb{I}$-firm located at r can pay. Then, as in Section 10.2.2, the agglomeration of the $\mathbb{I}$-firms in the city is a location equilibrium if and only if

$$\Omega(r) \leq 1 \quad \text{for all } r \geq 0.$$

To obtain $w^{\mathbb{I}}(r)$, we reformulate (10.13) within the present context. This means that the total expenditure on the $\mathbb{I}$-goods at the city place is equal to $\alpha c^{\mathbb{M}}(0)X_o^{\mathbb{M}}$, whereas that at $r \neq 0$ is equal to $\alpha c^{\mathbb{M}}(r)X^{\mathbb{M}}(r)$. Consequently, replacing $Y(0)$ by $c^{\mathbb{M}}(0)X_o^{\mathbb{M}}$, $Y(r)$ by $c^{\mathbb{M}}(r)X^{\mathbb{M}}(r)$, $\tau^{\mathbb{M}}$ by $\tau^{\mathbb{I}}$, and μ by α in (10.13) yields

$$w^{\mathbb{I}}(r) = \kappa_2' \Bigg\{ c^{\mathbb{M}}(0)X_o^{\mathbb{M}} \exp[-(\sigma-1)\tau^{\mathbb{I}}r][P^{\mathbb{I}}(0)]^{\sigma-1} + \int_{-r_b}^{r_b} c^{\mathbb{M}}(s)X^{\mathbb{M}}(s) \exp[-(\sigma-1)\tau^{\mathbb{I}}|r-s|][P^{\mathbb{I}}(s)]^{\sigma-1} ds \Bigg\}^{1/\sigma},$$

where $P^{\mathbb{I}}(r)$ is given by (10.32) and

$$\kappa_2' \equiv \rho[\alpha/(\sigma-1)f]^{1/\sigma}.$$

Furthermore, in the foregoing expression for $w^{\mathbb{I}}(r)$, we may set $c^{\mathbb{M}}(0)X_o^{\mathbb{M}} = L_o^{\mathbb{M}}/(1-\alpha)$, and $c^{\mathbb{M}}(r)X^{\mathbb{M}}(r) = w^*(r)L^{\mathbb{M}}(r)/(1-\alpha)$ by using (10.36) and (10.37) so that the potential function of an $\mathbb{I}$-firm (10.42) becomes

$$\Omega(r) = \frac{\alpha}{(1-\alpha)L^{\mathbb{I}}[w^*(r)]^{\sigma}} \Bigg\{ L_o^{\mathbb{M}} \exp[-(\sigma-1)\tau^{\mathbb{I}}r] + \int_{-r_b}^{r_b} w^*(s)L^{\mathbb{M}}(s) \exp[-(\sigma-1)\tau^{\mathbb{I}}(|r-s|-|s|)]\, ds \Bigg\}. \tag{10.43}$$

10.3.2 The Integrated City Equilibrium

10.3.2.1 Equilibrium

We consider here the configuration in which the entire production of both industrial sectors takes place within the city, thus implying that the city exports the $\mathbb{M}$-good and imports the $\mathbb{T}$-good. Hence, it must be that

$$X^{\mathbb{M}}(r) = 0 \text{ and } L^{\mathbb{M}}(r) = 0 \quad \text{for all } r \neq 0, \tag{10.44}$$

and thus the market-clearing condition for the $\mathbb{M}$-good can be written as follows:

$$X_o^{\mathbb{M}} = Q_o^{\mathbb{M}} + 2\int_0^{r_{\rm b}} Q^{\mathbb{M}}(r)\exp(\tau^{\mathbb{M}} r)\,dr.$$

Using (10.34), (10.35), and (10.44), we obtain

$$X_o^{\mathbb{M}} = \mu\frac{L_o^{\mathbb{M}} + L^{\mathbb{I}}}{p^{\mathbb{M}}(0)} + 2\int_0^{r_{\rm b}} \mu\frac{p^{\mathbb{T}}(r)}{p^{\mathbb{M}}(r)}\exp(\tau^{\mathbb{M}} r)\,dr. \tag{10.45}$$

Furthermore, to support that trade pattern, the equilibrium prices, $p^{\mathbb{T}}(r)$ and $p^{\mathbb{M}}(r)$, and the land rent, $R^*(r)$, must satisfy the relations (10.1), (10.2), and (10.3), respectively. Therefore, the wage at the agricultural border $r_{\rm b}$ is also given by (10.4).

Finally, the zero-profit condition for the $\mathbb{M}$-sector at the city place implies that

$$p^{\mathbb{M}}(0) = c^{\mathbb{M}}(0). \tag{10.46}$$

Using these conditions together with (10.29), (10.30), (10.32), (10.33), (10.39), and (10.40), we can easily solve the corresponding system for all the variables as functions of the single unknown $r_{\rm b}$. We thus obtain

$$L^{\mathbb{I}} = \alpha(L - 2c^{\mathbb{T}} r_{\rm b}) \tag{10.47}$$

$$L_o^{\mathbb{M}} = (1-\alpha)(L - 2c^{\mathbb{T}} r_{\rm b}) \tag{10.48}$$

$$p^{\mathbb{T}}(r) = c^{\mathbb{T}}[\exp\mu(\tau^{\mathbb{M}} + \tau^{\mathbb{I}})r_{\rm b}]\exp(-\tau^{\mathbb{T}} r) \tag{10.49}$$

$$p^{\mathbb{M}}(r) = \kappa_3\alpha^{-\alpha/(\sigma-1)}(L - 2c^{\mathbb{T}} r_{\rm b})^{-\alpha/(\sigma-1)}\exp(\tau^{\mathbb{M}} r) \tag{10.50}$$

$$w^*(r) = \exp[\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}}]r \tag{10.51}$$

$$P^{\mathbb{I}}(r) = \frac{1}{\rho}\left(\frac{\sigma f}{\alpha}\right)^{1/(\sigma-1)}(N - 2c^{\mathbb{T}} r_{\rm b})^{-1/(\sigma-1)}\exp(\tau^{\mathbb{I}} r). \tag{10.52}$$

To find $r_{\rm b}$, we substitute (10.36), (10.47), and (10.48) into the $\mathbb{M}$-good market-clearing condition (10.45) and use (10.1), (10.2), and (10.46). This leads to the relation

$$L - 2c^{\mathbb{T}} r_{\rm b} = \frac{2\mu c^{\mathbb{T}}}{1-\mu}\,\frac{1-\exp(-\tau^{\mathbb{T}} r_{\rm b})}{\tau^{\mathbb{T}}}\exp[\mu(\tau^{\mathbb{M}} + \tau^{\mathbb{T}})r_{\rm b}], \tag{10.53}$$

which is identical to (10.9).

Therefore, when the total population L is the same, both the monocentric economy equilibrium studied in Section 10.2 and the integrated city equilibrium considered here yield the same agricultural border ($r_{\rm b}$) and, hence, the same agricultural population ($L - 2c^{\mathbb{T}} r_{\rm b}$). The number of workers in the industry is therefore the same in both equilibria. It can also be readily verified that the equilibrium nominal wages, (10.11) and (10.51), are the same, as well as the equilibrium prices of the agricultural good. The reason for these seemingly surprising results is as follows. As long as the $\mathbb{M}$-sector and the $\mathbb{I}$-sector are agglomerated, the economy as a whole looks essentially like an economy with a single industrial sector involving increasing returns. Indeed, we have seen in Section 4.4 that, in the aggregate, the increasing returns appearing in the $\mathbb{I}$-sector are transferred to the $\mathbb{M}$-sector.

This analogy, however, does not carry over to all the microeconomic aspects of the economy. In particular, the equilibrium real wages are different. To obtain ω^* in the present context, we substitute (10.49) and (10.50) into (10.29) and use (10.52) as well as (10.53). We then find that

$$\omega^* = k_2[1 - \exp(-\tau^{\mathbb{T}} r_b)]^{\alpha\mu/(\sigma-1)} \exp\{[\alpha\mu/\rho - (1 - \mu + \alpha\mu)]\mu(\tau^{\mathbb{M}} + \tau^{\mathbb{T}})r_b\} \tag{10.54}$$

in which

$$k_2 \equiv \alpha^{\alpha\mu/\rho}(1-\alpha)^{(1-\alpha)\mu}\rho^{\alpha\mu}(c^{\mathbb{T}})^{-(1-\mu)}\left(\frac{2\mu c^{\mathbb{T}}}{(1-\mu)\sigma f \tau^{\mathbb{T}}}\right)^{\alpha\mu/(\sigma-1)},$$

which is indeed different from (10.10). Both expressions, however, have the same structure and are identical when $\alpha = 1$.[7]

Next, we turn to the sustainability of the integrated city as an equilibrium. First, we must check the equilibrium location conditions (10.41) for the $\mathbb{M}$-sector. Substituting (10.50) for $p^{\mathbb{M}}(r)$ while solving (10.33), (10.47), and (10.51) for $c^{\mathbb{M}}(r)$, we may rewrite Eq. (10.41) as follows:

$$\exp(\tau^{\mathbb{M}} r) \leq \{\exp(1-\alpha)[\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}}]r\}\exp(\alpha\tau^{\mathbb{I}} r)$$

or, equivalently, $(1-\alpha)(1-\mu)\tau^{\mathbb{T}} + (1-\mu+\alpha\mu)\tau^{\mathbb{M}} \leq \alpha\tau^{\mathbb{I}}$, namely,

$$\frac{(1-\alpha)(1-\mu)}{\alpha} + \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}.$$

Considering now the equilibrium location condition for the $\mathbb{I}$-firms, we can evaluate the potential function (10.43) by using (10.44), (10.47), (10.48), (10.51), (10.52), and (10.53) to obtain

$$\Omega(r) = \exp\{-\sigma[\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}} + \rho\tau^{\mathbb{I}}]r\}.$$

Hence, the equilibrium condition, $\Omega(r) \leq 1$, holds if and only if $\mu\tau^{\mathbb{M}} - (1-\mu)\tau^{\mathbb{T}} + \rho\tau^{\mathbb{I}} \geq 0$, or

$$\frac{1-\mu}{\rho} - \frac{\mu}{\rho}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}.$$

Therefore, we have shown the following:

Proposition 10.3 *The integrated city configuration is a spatial equilibrium if and only if the transport costs* $\tau^{\mathbb{T}}$, $\tau^{\mathbb{M}}$, *and* $\tau^{\mathbb{I}}$ *satisfy the following two conditions:*

$$\frac{(1-\alpha)(1-\mu)}{\alpha} + \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}} \tag{10.55}$$

and

$$\frac{1-\mu}{\rho} - \frac{\mu}{\rho}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}. \tag{10.56}$$

The parameter range in which these two conditions hold is described by the shaded area in Figure 10.3 for the case in which $(1-\alpha)/\alpha > 1/\rho$.[8]

The figure reveals that, for the integrated city to be an equilibrium, the transport costs of the intermediate commodities must be sufficiently high relative to those of the consumption goods.[9] This is because the vertical linkages between the two industrial sectors create a strong agglomeration force, a result that agrees with what we saw in Section 9.3. Yet, as shown by the wage function (10.51), any location in the countryside has a comparative advantage in labor cost for both sectors when $\mu\tau^{\mathbb{M}} < (1-\mu)\tau^{\mathbb{T}}$, that is, when shipping the $\mathbb{M}$-good is sufficiently cheap relative to the cost of shipping the agricultural good.

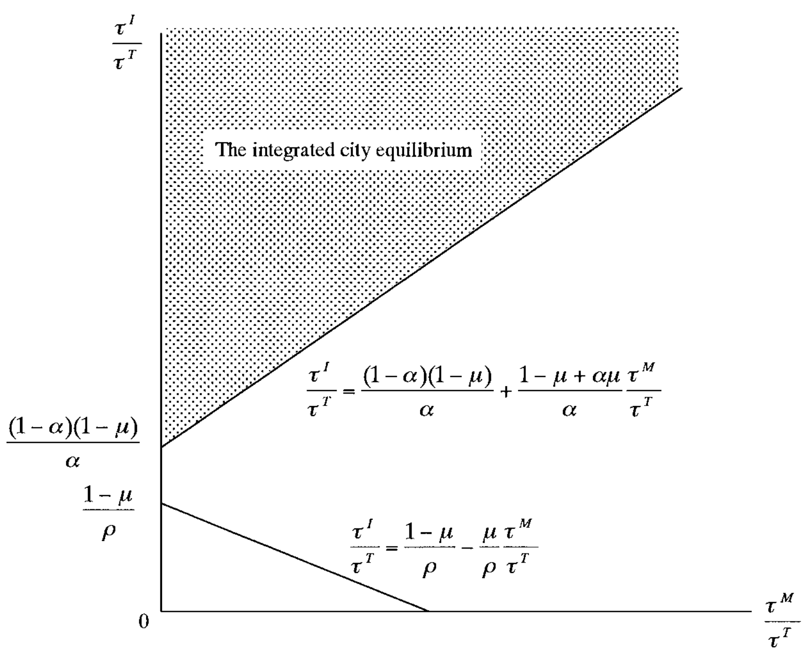


Figure 10.3: The parameter range for the integrated city equilibrium.

However, when shipping the $\mathbb{I}$-goods is costly enough, this advantage turns out to be more than outweighed by the fact $\mathbb{M}$-sector's having a high accessibility to its suppliers (the $\mathbb{I}$-firms), which are all located in the city, once it is itself concentrated within the city.

Observe also that conditions (10.55) and (10.56) are independent of L. As a result, increasing the population size does not affect the spatial pattern of the economy but results in a growing city and a larger agricultural area and never in the formation of new cities. Although the conditions stated in Proposition 10.3 differ from those given in Proposition 10.1, they can be viewed as the intermediate-good counterpart of the black hole condition (10.21) found in Section 10.2. However, their respective welfare implications differ.

10.3.2.2 Welfare

Our purpose is to study the impact of population growth on the welfare of economic agents. Clearly, the expression (10.24) relating landlords' real income and workers' real wage still holds at each location r. Because $p^{\mathbb{T}}(r)$ and $w^*(r)$ are unchanged, (10.24) and, therefore, (10.25) remain valid. Therefore, we have to sign the following expression derived from (10.54):

$$\frac{d\omega^*}{dr_b} = \frac{\mu\omega^*}{\sigma - 1}\left\{\sigma[\alpha\mu - (1 - \mu + \alpha\mu)\rho)](\tau^{\mathbb{M}} + \tau^{\mathbb{T}}) + \frac{\alpha\tau^{\mathbb{T}}}{\exp(\tau^{\mathbb{T}} r_b) - 1}\right\}, \tag{10.57}$$

which is identical to (10.23) when $\alpha = 1$. Using (10.25) and (10.57) as we did in Section 10.2.3, we can similarly show that

$$\frac{1}{\omega^L(r)}\frac{d\omega^L(r)}{dr_b} > \frac{\mu^2(\sigma + \alpha - 1)}{\sigma - 1}(\tau^{\mathbb{M}} + \tau^{\mathbb{T}}) > 0$$

regardless of the value of r_b. As in Proposition 10.2, we may then conclude as follows:

Proposition 10.4 *Consider the integrated city configuration. Then, if the population size increases*

1. *the welfare of the landlords always increases within the agricultural area;*
2. *when*

$$\rho \le \alpha\mu/(1 - \mu + \alpha\mu) \tag{10.58}$$

workers' welfare always increases;

3. *when*

$$\rho > \alpha\mu/(1 - \mu + \alpha\mu) \tag{10.59}$$

workers' welfare rises to some population level and declines beyond it.

It is worth noting that this proposition is very similar to Proposition 10.2 obtained in the case of a monocentric economy based on love for variety (they are even identical in the special case where $\alpha = 1$). By contrast, the sustainability conditions stated in Propositions 10.1 and 10.3 are very different. In particular, we know that raising the population size never destroys the integrated city equilibrium, although, beyond some threshold level, workers' welfare starts declining. In this case, the economy is in a situation of "primacy trap" in which workers' welfare declines with the growth of the metropolis whereas no other major city emerges. Today, some urban giants in developing countries (think of Bangkok or Jakarta) may be examples of such primate cities.

The primacy trap never arises, however, in the setting considered in the preceding section. On one hand, when $\rho \leq \mu$, the monocentric configuration remains an equilibrium when L increases, but the welfare of all agents also rises. On the other hand, when $\rho > \mu$, the welfare of workers declines beyond some population size; however, the monocentric configuration ceases to be an equilibrium when the population is sufficiently large, implying that new cities emerge and, hence, that workers' welfare starts growing again.

Such a difference in results occurs because, in the love for variety model of Section 10.2, the consumers themselves put together the varieties of the differentiated product, whereas here this is done by the $\mathbb{M}$-sector that sells a homogeneous product to consumers. In the former case, the producers of the differentiated varieties sell their output directly to the consumers, who are themselves dispersed, whereas, in the latter, the producers of the differentiated inputs sell their output to the $\mathbb{M}$-sector, which is entirely concentrated within the city. Hence, not surprisingly, the lock-in effect of an integrated city with intermediate commodities is much stronger than that of a monocentric economy based on love for variety.

10.3.3 The $\mathbb{I}$-Specialized City Equilibrium

10.3.3.1 Equilibrium

We now come to the case of the $\mathbb{I}$-specialized city in which only the intermediate sector is completely agglomerated within the city. By contrast, the production of the $\mathbb{M}$-good is decentralized across locations in a way such that each place is self-sufficient. Thus, the city exports the $\mathbb{I}$-varieties toward the hinterland, from which the $\mathbb{T}$-good is imported.

The self-sufficiency of each location in the $\mathbb{M}$-good implies that

$$X_o^{\mathbb{M}} = Q_o^{\mathbb{M}}$$
$$X^{\mathbb{M}}(r) = Q^{\mathbb{M}}(r) \quad 0 < r \leq r_{\text{b}}.$$

The equilibrium prices of the $\mathbb{M}$-good are implicitly determined by the zero-profit condition:

$$p^{\mathbb{M}}(r) = c^{\mathbb{M}}(r) \quad r \leq r_{\text{b}}.$$

For the agricultural good, the equilibrium price is given by (10.1), whereas the wage at the fringe of the inhabited area is (10.4).

As before, we may use these conditions together with those stated in Section 10.3.1 to determine all the variables as functions of the single unknown r_b:

$$L^{\mathbb{I}} = \frac{2\alpha\mu\tau^{\mathbb{T}}}{1-\mu}\frac{1-\exp(-\tau^{\mathbb{T}}r_b)}{\tau^{\mathbb{T}}}\exp[\alpha\mu(\tau^{\mathbb{I}}+\tau^{\mathbb{T}})/(1-\mu+\alpha\mu)]r_b \tag{10.60}$$

$$L_o^{\mathbb{M}} = \frac{\mu(1-\alpha)}{1-\mu+\alpha\mu}L^{\mathbb{I}} \tag{10.61}$$

$$L^{\mathbb{M}}(r) = \frac{\mu(1-\alpha)c^{\mathbb{T}}}{1-\mu+\alpha\mu}\exp[\alpha\mu(\tau^{\mathbb{I}}+\tau^{\mathbb{T}})/(1-\mu+\alpha\mu)](r_b-r) \quad r \le r_b \tag{10.62}$$

$$p^{\mathbb{T}}(r) = c^{\mathbb{T}}\exp[\alpha\mu(\tau^{\mathbb{I}}+\tau^{\mathbb{T}})/(1-\mu+\alpha\mu)]r_b\exp(-\tau^{\mathbb{T}}r) \tag{10.63}$$

$$p^{\mathbb{M}}(r) = \kappa_3(L^{\mathbb{I}})^{-\alpha/(\sigma-1)}\exp\{[\alpha\tau^{\mathbb{I}} - (1-\alpha)(1-\mu)\tau^{\mathbb{T}}]/(1-\mu+\alpha\mu)\}r \tag{10.64}$$

$$w^*(r) = \exp\{[\alpha\mu\tau^{\mathbb{I}} - (1-\mu)\tau^{\mathbb{T}}]/(1-\mu+\alpha\mu)\}r \tag{10.65}$$

$$P^{\mathbb{I}}(r) = \frac{1}{\rho}\left(\frac{\sigma f}{L^{\mathbb{I}}}\right)^{1/(\sigma-1)}\exp(\tau^{\mathbb{T}}r).$$

To obtain r_b, we substitute (10.61), (10.62), and (10.65) into the labor-market-clearing condition (10.40):

$$L - 2c^{\mathbb{T}}r_b = \frac{2\mu c^{\mathbb{T}}}{1-\mu+\alpha\mu}\left\{\frac{\alpha}{1-\mu}\frac{1-\exp(-\tau^{\mathbb{T}}r)}{\tau^{\mathbb{T}}} + \frac{(1-\alpha)[1-\exp(-\mathcal{K}r_b)]}{\mathcal{K}}\right\}\exp(\mathcal{K}r_b),$$

where

$$\mathcal{K} \equiv \alpha\mu(\tau^{\mathbb{I}}+\tau^{\mathbb{T}})/(1-\mu+\alpha\mu).$$

As in Section 10.2.2.2, it can readily be verified that r_b is uniquely determined and strictly increasing with L.

For the $\mathbb{I}$-specialized city configuration to be a spatial equilibrium, two additional conditions must be met. First, the assumption of no trade in the $\mathbb{M}$-good holds if and only if the rate of variation in the equilibrium prices of the $\mathbb{M}$-good never exceeds the transport cost $\tau^{\mathbb{M}}$ of this good. That is,

$$\frac{|dp^{\mathbb{M}}(r)/dr|}{p^{\mathbb{M}}(r)} \leq \tau^{\mathbb{M}} \quad \text{for all } r.$$

Given (10.64), this condition amounts to

$$\frac{|\alpha\tau^{\mathbb{I}} - (1-\alpha)(1-\mu)\tau^{\mathbb{T}}|}{1-\mu+\alpha\mu} \leq \tau^{\mathbb{M}}$$

or

$$\frac{(1-\alpha)(1-\mu)}{\alpha} - \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}$$
$$\leq \frac{(1-\alpha)(1-\mu)}{\alpha} + \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}}.$$

Second, the potential function of the $\mathbb{I}$-firms must never exceed 1. Substituting (10.60), (10.61), (10.62), and (10.65) into (10.43), we obtain the following expression for the potential function:

$$\Omega(r) = \frac{\alpha\mu}{1-\mu+\alpha\mu}\left\{\exp - \sigma\left[\frac{\alpha\mu\tau^{\mathbb{I}} - (1-\mu)\tau^{\mathbb{T}}}{1-\mu+\alpha\mu}\right]r\right\}$$
$$\times\left\{\exp[-(\sigma-1)\tau^{\mathbb{I}}r] + \frac{(1-\mu)\tau^{\mathbb{T}}}{2\alpha\mu(1-\exp(-\tau^{\mathbb{T}}r_{\mathrm{b}}))}\right.$$
$$\left.\times\int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} \exp[-\tau^{\mathbb{T}}|s| - (\sigma-1)\tau^{\mathbb{I}}(|r-s| - |s|)]ds\right\}.$$

Following the approach developed in the appendix, we may rewrite this expression as follows:

$$\Omega(r) = \exp(-\tilde{\eta}r)\left\{1 + \frac{(1-\mu)(\sigma-1)\tau^{\mathbb{I}}}{1-\mu+\alpha\mu}\int_0^r \exp[2(\sigma-1)\tau^{\mathbb{I}}s]\right.$$
$$\left.\times\left(1 - \frac{1-\exp(-\tau^{\mathbb{T}}s)}{1-\exp(-\tau^{\mathbb{T}}r_{\mathrm{b}})}\right)ds\right\}, \tag{10.66}$$

where

$$\tilde{\eta} \equiv \frac{\sigma\{[\alpha\mu + (1-\mu+\alpha\mu)\rho]\tau^{\mathbb{I}} - (1-\mu)\tau^{\mathbb{T}}\}}{1-\mu+\alpha\mu}.$$

Comparing (10.16) and (10.66), we see that the two expressions have essentially the same structure (when $\alpha = 1$ they are even identical). Accordingly, the location of the $\mathbb{I}$-firms may be studied as in Section 10.2.2.2. The following results are then obtained. First, because

$$\Omega'(0) = \sigma[(1-\mu)\tau^{\mathbb{T}} - \alpha\mu(1+\rho)\tau^{\mathbb{I}}]/(1-\mu+\alpha\mu)$$

a nonpositive slope at the origin requires that

$$\frac{1-\mu}{\alpha\mu(1+\rho)} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}.$$

Next, using the limit potential curve $\bar{\Omega}(r)$ associated with $r_{\mathrm{b}} \to \infty$, it can be shown that $\rho \leq \alpha\mu/(1-\mu+\alpha\mu)$ implies $\Omega(r) < 1$ for all r regardless of the positive value of r_{b}. By contrast, when $\rho > \alpha\mu/(1-\mu+\alpha\mu)$, the potential curves display the same patterns as those represented in Figure 10.2. This means that a critical value $\hat{L}$ exists such that the potential function is just equal to 1 at a particular location $\hat{r}$. Hence, for any value of L exceeding $\hat{L}$, $\Omega(r) > 1$ for a sufficiently large r and the specialized city configuration ceases to be a spatial equilibrium. Therefore, putting all these results together, we have:

Proposition 10.5 *Consider an $\mathbb{I}$-specialized city configuration. For this configuration to be a spatial equilibrium, it is necessary that*

$$\frac{(1-\alpha)(1-\mu)}{\alpha} - \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}} \leq \frac{(1-\alpha)(1-\mu)}{\alpha} + \frac{1-\mu+\alpha\mu}{\alpha}\frac{\tau^{\mathbb{M}}}{\tau^{\mathbb{T}}} \tag{10.67}$$

and

$$\frac{1-\mu}{\alpha\mu(1+\rho)} \leq \frac{\tau^{\mathbb{I}}}{\tau^{\mathbb{T}}}. \tag{10.68}$$

1. *If the foregoing conditions hold and if*

$$\rho \leq \alpha\mu/(1-\mu+\alpha\mu), \tag{10.69}$$

then the specialized city configuration is an equilibrium regardless of the population size.

2. *If the foregoing conditions hold and if*

$$\rho > \alpha\mu/(1-\mu+\alpha\mu), \tag{10.70}$$

then a critical population level $\hat{L}$ exists such that the specialized city configuration is an equilibrium for all $L \leq \hat{L}$, whereas it is no longer an equilibrium as soon as $L > \hat{L}$.

In Figure 10.4, we represent the domains of transport cost values for which the two equilibrium conditions hold.

This figure reveals that the $\mathbb{I}$-specialized city configuration is a spatial equilibrium when the transport cost of the $\mathbb{I}$-commodities is sufficiently low when compared with that of the $\mathbb{M}$-good. However, it should not be too low either, for otherwise $\mathbb{I}$-firms would move into the hinterland, where they can benefit from the lower wages prevailing there. But this is not yet the end of the story. When

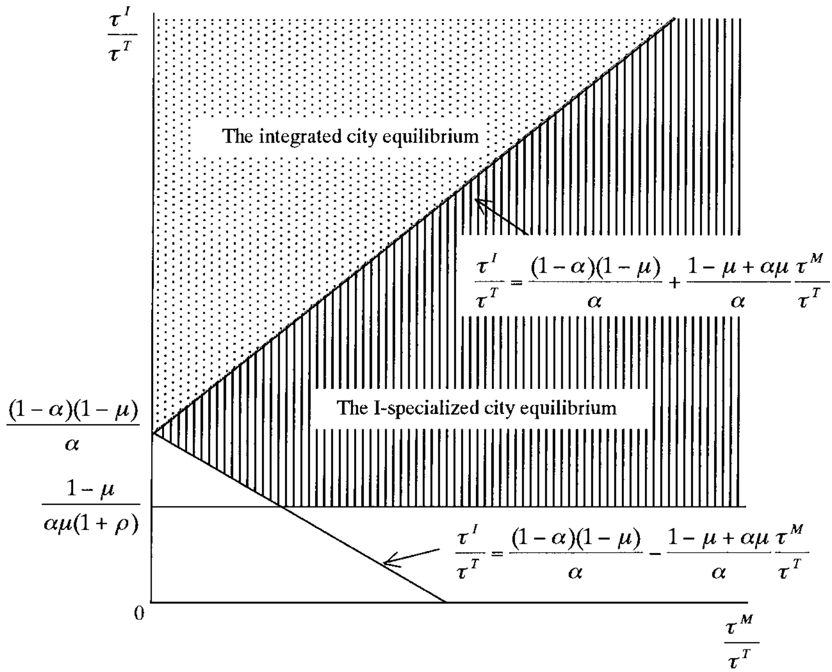


Figure 10.4: The parameter ranges for the integrated and $\mathbb{I}$-specialized city equilibria.

the intermediate inputs are good substitutes (that is, (10.70) holds), then this configuration is an equilibrium only when the population does not become too large. Thus, the $\mathbb{I}$-specialized city displays an equilibrium pattern very similar to that of the monocentric economy described in Proposition 10.1. In particular, the condition (10.69) is the new black hole condition, which reduces to the previous one, $\rho \leq \mu$, when $\alpha = 1$.

10.3.3.2 Welfare

We again study the impact of population growth on workers' and landlords' welfare. To this end, we first substitute (10.63) and (10.64) into (10.29) and then use (10.60) to obtain the equilibrium real wage:

$$\omega^* = k_2[1 - \exp(-\tau^{\mathbb{T}} r_b)]^{\alpha\mu/(\sigma-1)} \times \exp\left[\frac{\alpha\mu/\rho - (1 - \mu + \alpha\mu)}{1 - \mu + \alpha\mu}\alpha\mu(\tau^{\mathbb{I}} + \tau^{\mathbb{T}})r_b\right]. \tag{10.71}$$

Furthermore, because (10.24) still holds, substituting (10.63) and (10.64) into (10.24) and using (10.60), we obtain the equilibrium real income of the landlords:

$$\omega^L(r) = \omega^* c^{\mathbb{T}} \left\{ \exp\left[\frac{\alpha\mu(\tau^{\mathbb{I}} + \tau^{\mathbb{T}})}{1 - \mu + \alpha\mu}(r_b - r) \right] - 1 \right\}.$$

Differentiating (10.71) with respect to r_b yields

$$\frac{d\omega^*}{dr_b} = \frac{\mu\omega^*}{\sigma - 1} \left\{ \frac{\alpha}{1 - \mu + \alpha\mu} \sigma[\alpha\mu - (1 - \mu + \alpha\mu)\rho](\tau^{\mathbb{M}} + \tau^{\mathbb{T}}) + \frac{\alpha\tau^{\mathbb{T}}}{\exp\left(\tau^{\mathbb{T}} r_b\right) - 1} \right\}.$$

Using the same approach as in Section 10.3.2.2, we also obtain

$$\frac{1}{\omega^L(r)} \frac{d\omega^L(r)}{dr_b} > \frac{\alpha}{1 - \mu + \alpha\mu} \frac{\mu^2(\sigma + \alpha - 1)}{\sigma - 1}(\tau^{\mathbb{M}} + \tau^{\mathbb{T}}) > 0$$

regardless of the value of r_b.

These two expressions are similar to those given in Section 10.3.2.2. In fact, because $1 - \mu + \alpha\mu > 0$, we may conclude that Proposition 10.4 also holds for the I-specialized city configuration; however, the two critical population sizes beyond which workers' welfare declines are not necessarily the same.

Although the impact of population growth seems to be the same in both equilibria, there is a substantial difference that is worth noticing. If the economy is in an integrated city equilibrium, when $\rho \leq \alpha\mu/(1 - \mu + \alpha\mu)$, workers' welfare keeps declining once the population is above some critical value. By contrast, in an $\mathbb{I}$-specialized city equilibrium, when $\rho \leq \alpha\mu/(1 - \mu + \alpha\mu)$ the growth of population eventually leads to the formation of new cities, thus boosting the welfare of workers. This suggests a possible strategy for the economy to escape from the primacy trap. Indeed, inspecting Figure 10.4 shows that a policy allowing for a reduction in the transport costs of the intermediate commodities should move the economy into the domain for which the economy is in an $\mathbb{I}$-specialized city equilibrium.[10] The growth of population is then accompanied by the formation of new cities and, therefore, an increase in workers' welfare.

10.4 ON THE EMERGENCE AND STRUCTURE OF URBAN SYSTEMS

When the city is not a black hole, the monocentric configuration ceases to be a spatial equilibrium once the population size becomes sufficiently large. A glance at Figure 10.2 reveals that, when the population reaches the level $\hat{L}$, the potential curve just hits the value 1 at location $\hat{r}$. Even though no agglomeration exists there, this location becomes as attractive as the incumbent city because firms now have a direct access to a large local market situated deep inside the hinterland. This suggests that the relocation of an arbitrary small (but positive) mass of firms at $\hat{r}$ is able to trigger a mechanism of agglomeration that leads to the creation of a new city. Simultaneously, a mechanism of contraction is at work in the existing city, but the lock-in effect prevents this city from disappearing. Hence, the new city does not capture the whole population of the incumbent city. By symmetry, the same arises for $r < 0$, and thus the economy actually moves from a one-city configuration to a three-city system. However, for the moment, we focus on the area $r > 0$.

When the population keeps rising, the agricultural frontier moves farther and farther away from the new city because a growing population must be supported. Accordingly, for the same reason as before, when the population reaches some threshold level, a potential curve similar to the original one hits the value 1 at a new location. This leads to the emergence of an additional city there. With a growing population and an unbounded space, a new city will be created periodically at a certain distance from the nearest existing city. In this way, a system of cities is formed in which cities are located at (more or less) equal distances.

In this section, we use the framework of Section 10.2 to make the foregoing ideas more precise. When the monocentric pattern described in Proposition 10.1 ceases to be a stable equilibrium, the economy moves to a new stable equilibrium involving the formation of a new city. This is achieved through the decisions made by a myriad of agents motivated only by their own interests. The contrast with the approach taken in Chapters 4 and 5 is, therefore, striking.

However, to study stability, we must describe the behavior of economic agents when the economy is not at equilibrium. Our first step is, therefore, to describe how the transition works, using an adjustment process followed by the agents. We then use this process to show how the combination of a growing population and of a homogeneous space leads to the formation of a regular network of cities. Because the formal analysis is long and complex and details can be found in the cited references, we will refrain from trying to be complete and will restrict ourselves to the presentation of the main ideas.

10.4.1 The Adjustment Process

We assume, again, that the total population grows slowly. Our purpose is to study how the spatial economy evolves as a consequence of this growth. As discussed above, changes in the spatial configuration take the form of new cities, which appear once the existing spatial system becomes unstable. To assess stability, we must specify an adjustment process. The spirit of this process is similar to the one used in Chapter 9, which is driven by workers' migration toward locations offering higher real wages. There are, however, several differences. First, the set of potential locations for new cities is now infinite, thus making the set of possible spatial configurations much larger than before. Furthermore, as the population grows, the economy entails a sequence of changes in the spatial distribution of industry that correspond to different spatial equilibria (by contrast, we focused on a single transition in Chapter 9).

Our research strategy is to use "structural stability" of spatial equilibria as a selection device at each time t. Accordingly, when studying the stability of the spatial equilibrium prevailing at any given time t, we first assume that the population $L(t)$ is momentarily fixed and then analyze the impact of small perturbations of the population distribution equilibrium over what we call "fictitious time" denoted by ξ. To this end, consider a set with any number K of sites, which includes the actual cities, whereas the remaining sites represent potential locations for new cities. During the adjustment process, the population of the kth city is denoted by $L_k(\xi) \geq 0$. At the fixed time t, the workers' equation of motion over ξ is as follows:

$$\frac{dL_k(\xi)}{d\xi} = L_k(\xi)[\omega_k(\xi) - \bar{\omega}(\xi)]L(t) \quad k = 1, \ldots, K \tag{10.72}$$

in which $\omega_k(\xi)$ is the temporary-equilibrium real wage prevailing in city k, whereas $\bar{\omega}(\xi)$ is the average real wage across all workers and farmers in the economy:

$$\bar{\omega}(\xi) = \left\{ \sum_{k=1}^{K} L_k(\xi)\omega_k(\xi) + \left[L(t) - \sum_{k=1}^{K} L_k(\xi) \right] \omega^{\mathbb{T}}(\xi) \right\} \Big/ L(t),$$

where $\omega^{\mathbb{T}}(\xi)$ is the temporary-equilibrium real wage common to all agricultural workers. Because the total population of workers and farmers is equal to $L(t)$, the mass of farmers, $L^{\mathbb{T}}(\xi)$, must obey the following equation:

$$\frac{dL^{\mathbb{T}}(\xi)}{d\xi} = L^{\mathbb{T}}(\xi)[\omega^{\mathbb{T}}(\xi) - \bar{\omega}(\xi)]L(t). \tag{10.73}$$

Equations (10.72) and (10.73) describe the migration of workers and farmers across the whole inhabited area; this area is itself variable because the fringe moves with ξ.

Stability of a spatial equilibrium at time t requires the following condition: for any number K (such that K is never lower than the number of actual cities in equilibrium) of sites at any possible locations (but in which the locations of actual cities are unchanged), *the spatial equilibrium is stable under the system (10.72) and (10.73).*[11] This definition is proposed to accommodate the infinite number of possible city locations.

At first sight, checking stability looks like a formidable task. However, given a spatial equilibrium, it can be shown that if the associated potential curve is strictly less than 1 but at the locations of existing cities, the spatial system is stable, thus implying that no new city is viable at any location (Fujita and Mori 1995; 1997). Intuitively, this can be understood as follows. When the economy is in equilibrium, workers' real wage in each city and farmers' real wage are equal across all locations. Because

$$\Omega(r) = \left(\frac{w^{\mathbb{M}}(r)}{w^*(r)}\right)^{\sigma} = \left(\frac{\omega^{\mathbb{M}}(r)}{\omega^*(r)}\right)^{\sigma}$$

and $\Omega(r) < 1$ at any r that does not accommodate a city, there is no place, other than the incumbent cities, at which firms can offer workers the prevailing equilibrium real wage while making nonnegative profits.

10.4.2 The Urban System as a Network of Cities

Consider an economy having a monocentric configuration associated with a small population (so that the potential curve is lower than 1 for all $r \neq 0$). In this case, the equilibrium configuration is stable and new cities cannot emerge. However, as seen in Section 10.2, increasing the population size shifts the potential curve up. Eventually, this curve hits the value 1 at some location $\hat{r}$ when the population size reaches the level $\hat{L}$, thus making this location as profitable as the city. If the population slightly rises above $\hat{L}$, then the potential curve exceeds 1 at $\hat{r}$. Hence, $\hat{r}$ is now more profitable than the city, which leads some firms to set up there. In other words, a new city is forming at $\hat{r}$. Of course, the same arises at location $-\hat{r}$. Thus, at the population level $\hat{L}$, the monocentric equilibrium is transformed into a symmetric tricentric pattern.[12] If the multiplier effect encapsulated in the agglomeration mechanism is sufficiently powerful, the resulting change will be catastrophic, resulting in the emergence of two fairly large flanking cities. In this respect, Fujita et al. (1999, 168) have shown that sufficient conditions for such a catastrophic transition are as follows:

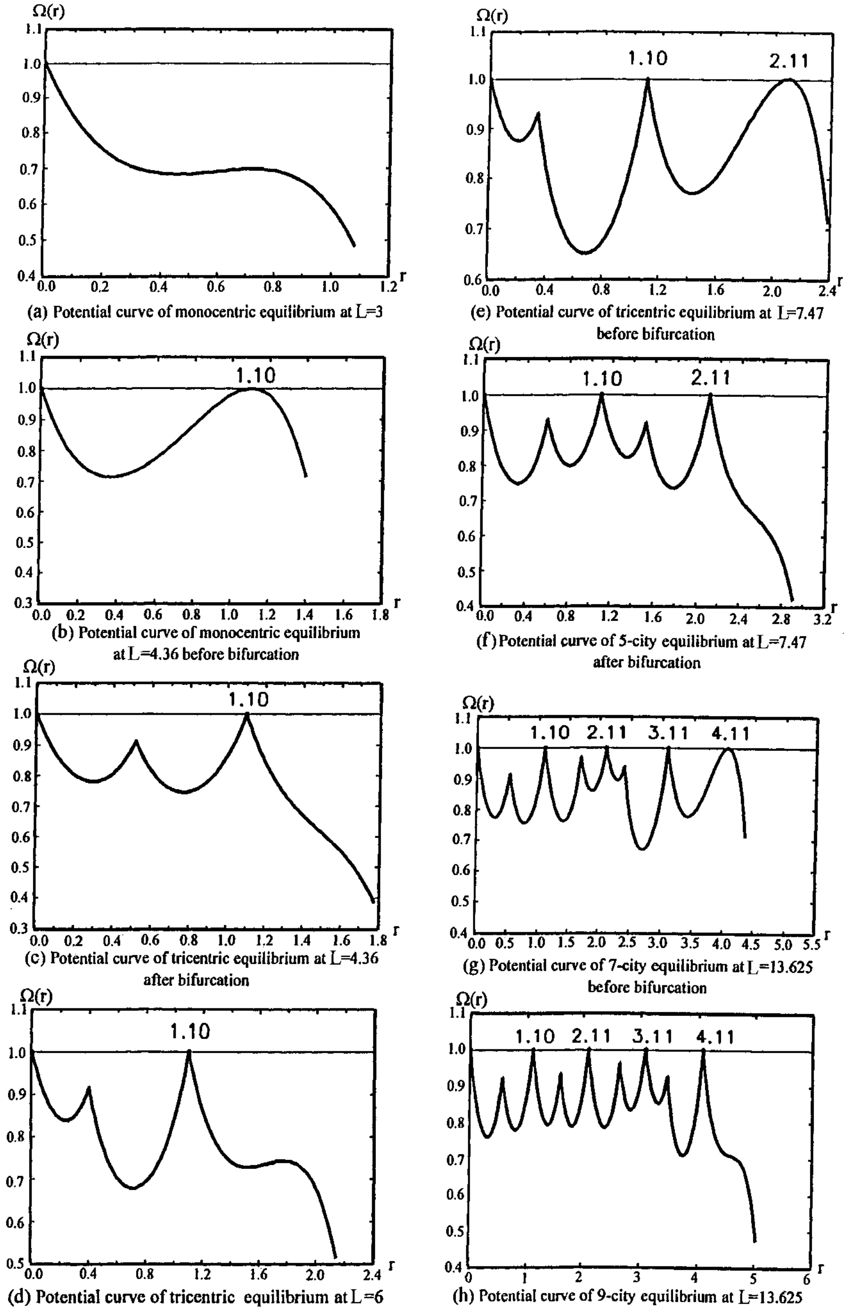
$$\mu\tau^{\mathbb{M}} \geq (1-\mu)\tau^{\mathbb{T}} \tag{10.74}$$

and

$$\mu \geq \frac{\rho}{1+\rho}. \tag{10.75}$$

Equation (10.74) means that transport costs of the industrial goods, weighted by their expenditure share, are to be large enough when compared with those of the agricultural goods, and thus being located in a city allows consumers to raise their real income markedly. Equation (10.75) states that the share of the industrial goods in consumption must be above some threshold that depends negatively on the degree of product differentiation. Indeed, a weak share in consumption would not provide enough incentive for firms and consumers to form a large agglomeration. Of course, μ cannot be larger than ρ, for otherwise the agglomeration force is so strong that the incumbent city acts as a black hole and no new city can emerge.

It appears to be very problematic to derive analytical results when the equilibrium involves more than three cities. We thus restrict ourselves to a discussion of the numerical results obtained by Fujita and Mori (1997). The parameters of



(a) Potential curve of monocentric equilibrium at L=3

(b) Potential curve of monocentric equilibrium at L=4.36 before bifurcation

(c) Potential curve of tricentric equilibrium at L=4.36 after bifurcation

(d) Potential curve of tricentric equilibrium at L=6

(e) Potential curve of tricentric equilibrium at L=7.47 before bifurcation

(f) Potential curve of 5-city equilibrium at L=7.47 after bifurcation

(g) Potential curve of 7-city equilibrium at L=13.625 before bifurcation

(h) Potential curve of 9-city equilibrium at L=13.625

Figure 10.5: Evolution of an urban system in a spatial economy.

the economy take the values corresponding to Figure 10.2; they clearly satisfy (10.74) and (10.75).

Figure 10.5 describes the evolution of the urban system as L keeps rising. Figure 10.5 (a), drawn when $L = 3$, shows that the monocentric configuration is a stable equilibrium. In Figure 10.5 (b), drawn when L takes the critical value 4.36, the monocentric configuration becomes structurally unstable because the corresponding potential curve hits the value 1 at $r = 1.10$. At this particular value of L, the spatial economy experiences a catastrophic transition from a monocentric to a tricentric pattern. As shown by the corresponding potential curve depicted in Figure 10.5 (c), this new equilibrium is stable.

In Figure 10.5 (d), the tricentric equilibrium is shown to be stable for $L = 6$, that is, a value of L that lies between the first and second bifurcations. In Figure 10.5 (e), L takes the new critical value 7.47 and the corresponding potential curve hits the value 1 at $r = 2.11$, which indicates that the tricentric pattern becomes structurally unstable. Again, there is a catastrophic bifurcation such that the tricentric pattern is transformed into a pentacentric configuration. Figure 10.5 (f) describes the potential curve after the transition.

In a similar manner, as L continues to rise, a pair of frontier cities emerge periodically as a result of a catastrophic transition of the spatial economy. Figures 10.5 (g) and (h) describe another example of such a bifurcation from a heptacentric to a nonacentric configuration in which the new city arises at $r = 4.11$.

These diagrams suggest that, as the number of cities increases, the urban system approaches a highly regular network of cities, as conjectured in central place theory.

10.5 CONCLUDING REMARKS

We have seen in this chapter that the addition of a land market to the canonical model of Chapter 9, together with the assumption of mobile workers, allows for the study of a process of urbanization in which firms balance the advantages of being close to concentrated output markets against those associated with a low degree of competition in rural areas. Despite severe limitations, such a process captures some of the basic forces stressed by geographers and historians. Not surprisingly, it appears that the number of cities expands with the size of the total population. Our emphasis on population growth as a major factor for urbanization agrees with economic history as well as with the observations made by early scholars such as A. Smith ([1776] 1965, 17), who observed that

> in the lone houses and very small villages which are scattered about in so desert a country as the Highlands of Scotland, every farmer must be butcher, baker and brewer for his own family.

Stated differently, population must become dense enough for the division of labor to expand through the emergence of new cities. The finer division of labor is expressed here through a larger variety of goods available to consumers.

More surprisingly perhaps, new cities distribute themselves according to a (fairly) regular network in which differentiated varieties are traded. This is in accordance with some of the main principles of central place theory. It is worth noting, however, that if cities serve as a basis for the agricultural area, they are also engaged in trade because they are specialized in the production of different varieties. This implies that *a city has a twofold function, namely serving its agricultural hinterland and trading with other cities,* as suggested by the work of many economic historians (see, e.g., Hohenberg and Lees 1985, chap. 2).

Using a framework similar to ours, Mori (1997) investigated the impact of a decrease in the cost of shipping the industrial goods and obtained an interesting set of results. Some firms producing the $\mathbb{M}$-goods choose to locate away from cities because (1) lower wages are paid in the agricultural areas and (2) the supplying costs of the $\mathbb{M}$-goods to cities are low. Likewise, workers are willing to leave cities because the cost of agricultural goods is lower; therefore, real wages are higher in the rural hinterlands, whereas the $\mathbb{M}$-goods are available at prices close to those prevailing in cities. Although new cities may emerge when the population is large enough, a substantial decrease in the cost of transporting the industrial goods fosters a more dispersed pattern of production in the industrial sector, which is a scheme resembling industrial belts observed in the real world.

However, only one type of city emerges in this urbanization process. Thus, the model considered fails to grasp the fundamental aspect of an urban hierarchy. It is our contention, however, that the approach developed in this chapter may be enriched to address this difficult question. A first step in this direction was taken by Fujita, Krugman, and Mori (1999), who introduced several $\mathbb{M}$-goods into the utility (9.1), each having different elasticities of substitution (σ_i), transportation costs ($\tau_i^{\mathbb{M}}$), or both. They assumed that trade costs are equal across goods and that

$$\sum_{k=1}^{K} \frac{\mu_k}{\rho_k} \geq 1,$$

which is a condition that boils down to the black hole condition $\mu \geq \rho$ when $K = 1$ (see Section 9.2.3). As the population size increases, these authors have shown that a (more or less) regular hierarchical central place system, reminiscent of Christaller's, emerges within the economy. In this urban system, "higher-order cities" provide a larger number of groups of $\mathbb{M}$-goods. However, there is two-way trade between cities because cities are specialized in the production of differentiated goods. This leads to a more intricate pattern of trade in which horizontal relations are superimposed on the pyramidal structure of central place theory. As expected, higher-order cities export more varieties than lower-order cities. However, horizontal relations between cities of the same order may be more important than trade with lower-order cities. Thus, the urban hierarchy obtained in this way is more fuzzy than in the Christaller model of central places, although this hierarchy seems to reflect the urban systems of advanced economies well, as studied by Pred (1977).

APPENDIX

We may rewrite (10.15) as follows:

$$\Omega(r) = \mu A(r) \exp(-\eta r), \tag{A.1}$$

where η is defined by (10.17), whereas

$$\begin{aligned} A(r) &\equiv 1 + \frac{1-\mu}{2\mu} \frac{\tau^{\mathbb{T}} \exp[(\sigma-1)\tau^{\mathbb{M}} r]}{1-\exp(-\tau^{\mathbb{T}} r_{\mathrm{b}})} \\ &\quad \cdot \int_{-r_{\mathrm{b}}}^{r_{\mathrm{b}}} \exp[-\tau^{\mathbb{T}}|s| - (\sigma-1)\tau^{\mathbb{M}}(|r-s| - |s|)]\, ds \\ &= 1 + \frac{1-\mu}{2\mu} \frac{\tau^{\mathbb{T}}}{1-\exp\left(-\tau^{\mathbb{T}} r_{\mathrm{b}}\right)} \left\{ \int_{-r_{\mathrm{b}}}^{r} \exp[-\tau^{\mathbb{T}}|s| \right. \\ &\quad \left. + (\sigma-1)\tau^{\mathbb{M}}(s+|s|)]ds + \exp[2(\sigma-1)\tau^{\mathbb{M}} r] \int_{r}^{r_{\mathrm{b}}} \exp(-\tau^{\mathbb{T}} s)\, ds \right\}. \end{aligned}$$

Taking the logarithm of (A.1) and differentiating both sides of the resulting expression with respect to r, we obtain

$$\frac{\Omega'(r)}{\Omega(r)} = -\eta + \frac{A'(r)}{A(r)},$$

or, using (A.1), we get

$$\Omega'(r) = -\eta\Omega(r) + \mu \exp(-\eta r) A'(r), \tag{A.2}$$

where

$$A'(r) = \frac{1-\mu}{\mu} \frac{(\sigma-1)\tau^{\mathbb{M}}}{1-\exp(-\tau^{\mathbb{T}} r_{\mathrm{b}})} \exp[2(\sigma-1)\tau^{\mathbb{M}} r] \int_{r}^{r_{\mathrm{b}}} \exp(-\tau^{\mathbb{T}} s)\, ds.$$

Solving the differential equation (A.2) under the initial condition $\Omega(0) = 1$ leads to the expression

$$\Omega(r) = \exp(-\eta r)\left[1 + \mu \int_0^r A'(s)\, ds\right],$$

which is identical to (10.16).

NOTES

1. What Krugman (1995, chap. 2) has called the "Germanic geometry."
2. Once more, this idea has been anticipated by Thünen: "For instance, a countryman may visit the capital to sell his products, and decide to buy some liquor. It will be cheaper for him to buy this in the capital, even if it costs him half a thaler more than he would pay in the provincial town two miles from his farm, because he would have to make a special journey to fetch the local alcohol" (p. 287 of the English translation).
3. See Henderson (1972) for an early critical economic evaluation of central place theory.
4. As discussed in Section 1.3.3, Thünen himself provided an amazingly comprehensive theory on industrial agglomeration. However, not surprisingly, he was unable to develop a unified model of the isolated state in which his agricultural land use theory would be combined with his pioneering work on industrial location. This is because such a unified model requires a spatial general equilibrium model based on increasing returns.
5. See Becker and Henderson (2000) for a first comparison of the two modeling strategies on urbanization.
6. We will see in Section 10.4 that the spatial equilibrium identified in the proposition is also stable in a sense that will be explained there.
7. To show this, it is sufficient to set $0^0 = 1$ in (10.54), which is then identical to (10.10). Indeed, when $\alpha = 1$ in the production (Eq. (34) of Chapter 9), it does not make any difference in the aggregate whether the differentiated varieties are directly assembled by the consumers or through the perfectly competitive $\mathbb{M}$-sector.
8. When the opposite inequality holds, the two intercepts along the vertical axis are reversed. This difference, however, is not important for our discussion of the equilibrium.
9. In fact, several models of urban agglomeration based on intermediate commodities assume that these goods are nontradable (see, e.g., Rivera-Batiz 1988 and Abdel-Rahman and Fujita 1990). Thus, they can be considered as special cases of the present model.
10. This could be achieved by the construction of modern telecommunication infrastructure and high-speed passenger transport systems.
11. When K is fixed, this definition is equivalent to the standard one: When the spatial distribution of workers experiences any arbitrarily small perturbation away from the equilibrium, the state of the economy is stable if the spatial distribution always goes back to what it was; if it does not, the equilibrium is unstable.
12. Note that asymmetric configurations may also arise. We concentrate here on symmetric configurations because they convey the main message of the model.

11

On the Relationship between Agglomeration and Growth

11.1 INTRODUCTION

The likely main effect of interregional and international integration is to increase economic efficiency within the space-economy. However, as seen in Chapter 9, market expansion may well be accompanied by the development of some core regions whose wealth is, in part, obtained at the expense of peripheral regions: the average welfare in the region accommodating the modern sector rises but decreases in the other region. So far, however, such a result has been obtained using static models in which the total number of firms and varieties, which is determined by the parameters of the economy, is constant. It is, therefore, of fundamental importance to determine what the core–periphery property becomes in a dynamic setting in which the number of firms grows over time. In other words, we want to deal with the following question: How do growth and location affect each other? More precisely, we want to know whether regional discrepancies widen or fall over time, as well as the main reasons for such a possible divergence or convergence. Because in the context of human history, the current disparities between rich and poor regions are recent (Bairoch 1993, chap. 9), it is important to understand how they may change over time. Regional discrepancies are often considered socially undesirable, and the issue is consequently critical from the policy standpoint. If compelling evidence is found that the existing disparities are likely to persist or, worse, that growth and agglomeration will make those living in the periphery worse off, then governments and international bodies should become more active in designing policies to foster a more equitable distribution of wealth across nations and regions.

It has long been argued that growth is localized, for technological and social innovations tend to be spatially clustered whereas their diffusion across places is slow. For example, Hirschman (1958, 183) claimed that[1]

> we may take it for granted that economic progress does not appear everywhere at the same time and that once it has appeared powerful forces make for a spatial concentration of economic growth around the initial starting points.

Myrdal (1957, 26) similarly argued that

> the main idea I want to convey is that the play of the forces in the market normally tends to increase, rather than to decrease, the inequalities between regions.

More recently, Feldman and Florida (1994) observed that in the late twentieth century innovations clustered geographically in areas where research and development (R&D)-oriented firms or universities were established. These authors also observed that concentrations of such specialized resources reinforce a region's capacity to innovate and to grow. In this way, the connection between growth and geography becomes even stronger when regional specialization in innovation activity is viewed as the outcome of combining specific capabilities and capacities developed in those regions. Such an approach to innovation suggests that the process of development is similar to the one we have encountered in the formation of regional agglomerations. As a result, agglomeration can be considered the territorial counterpart of economic growth.

Connections between growth and cities are also under scrutiny in the modern literature on growth and development. The role of cities in economic growth since the second half of the nineteenth century has been emphasized by economic historians (Hohenberg and Lees 1985, chaps. 6 and 7). More precisely, cities are viewed as the main social institutions in which technological and social innovations are developed through market and nonmarket interactions. Furthermore, city specialization changes over time, thus creating a geographically diversified pattern of economic development.[2] For all these reasons, cities are often considered the engines of growth.

Clearly, space and time are intrinsically mixed in the process of economic development. However, the study of their interaction is a formidable task. Because either agglomeration or growth is a complex phenomenon by itself, one should expect any integrated analysis to face many conceptual and analytical hurdles. Not surprisingly, therefore, the field is still in its infancy, and relevant contributions have been few. Even so, however, it is beyond the reach of this chapter to provide a complete survey of the field of regional growth. Instead, we have chosen to present some results that are strongly linked to previous chapters in the hope of shedding light on the main factors in play and encouraging future research in this important domain. More precisely, we will see that there are different modeling strategies that rest mainly on the assumptions made regarding, first, the mobility of skilled labor and, second, on the intensity of spillover effects across regions. How to discriminate between the corresponding models is an empirical question that is not addressed here.

Moreover, the introduction of workers' migration into an endogenous growth model under perfect foresight raises unanticipated problems. Despite these difficulties, it is possible to derive some tentative conclusions that appear to be reasonable. Because both the "new" theories of growth and "new economic geography" share the same basic framework of monopolistic competition, a solid foundation for cross-fertilization exists between the two fields. And, indeed, the few existing contributions that have recently explored the mutual influences between growth and location exploit this formal analogy; see, for example, Waltz (1996), Baldwin (1999), Martin and Ottaviano (1999; 2001), and Baldwin, Martin, and Ottaviano (2001).

In accord with what we argued in previous chapters, we assume here that skilled people working for the R&D sector are mobile. To the best of our knowledge, today only three articles allow for labor mobility in a multiregional (or multicity) endogenous growth model under perfect foresight, that is, Waltz (1996), Baldwin and Forslid (2000), and Black and Henderson (1999). Although they represent pioneering works toward a unified analysis of growth and location, their treatment of migration seems to leave room for further consideration. The assumption of costless migration in Waltz (1996) results in a bang-bang migration behavior that is too far from reality. Although the article by Baldwin and Forslid (2000) represents one of the first attempts to merge the core–periphery literature with endogenous growth theory, its implications are not clear owing to analytical complexity. Finally, the endogenous growth model of an urban system by Black and Henderson (1999) represents a significant achievement; however, its implicit assumption of an optimally controlled migration process seems fairly restrictive.

In Section 11.2, we propose a simple model of endogenous growth for a two-region economy that represents a natural combination of a Krugman-type core–periphery model and a Grossman–Helpman–Romer-type model of endogenous growth with horizontally differentiated products. Specifically, this means that we add an R&D sector to the core–periphery model that uses skilled labor to create new varieties for the modern sector so that the number of firms producing in this sector is variable. In addition, forward-looking behavior and migration are formalized in the same spirit as in Section 9.5.[3] This model may be viewed as an attempt at integrating several issues addressed in previous chapters. More precisely, it combines (1) the demand effect generated by the migration of skilled workers as in Chapter 9 and (2) the productivity effect generated by the existence of spillovers, such as those studied in Chapter 6. These two effects are in turn associated with the growth in the number of varieties, which gives rise to a second demand effect. Hence, this model provides a "general" framework of reference that is amenable to a complete analytical solution.

For simplicity, we neglect the transitional period (except in the stability analysis) and focus on a steady-state spatial equilibrium; this equilibrium is such that the spatial distribution of skilled workers is time invariant, whereas the total number of varieties, and that of firm grow at a constant rate, both being determined at the equilibrium outcome. One of the most stimulating results obtained in this chapter is that the growth rate, measured by the variation in the number of varieties, changes with the spatial distribution of skilled workers. In other words, we show that *the growth of the global economy depends on the spatial organization of the innovation sector across regions*.

In Section 11.3, we assume that patents for new products can be transferred costlessly between regions. In this case, agglomeration economies turn out to be so strong that the entire R&D activity always concentrates into a single region. In addition, the modern sector is either fully or partially agglomerated in the same region as the R&D sector. In Section 11.4, we move to the other polar case in which patents developed in one region are not transferable to the other – presumably because there are social and cultural barriers to the adoption of new technologies. For example, it is well known that problems hamper the effective implementation of blueprints in a foreign region such as the tacit knowledge they require, which is hard to transfer abroad (Teece 1977). In such a context, we obtain results that are essentially similar to those of the static core–periphery model of Chapter 9. More precisely, the core–periphery structure in which the innovation and the modern sectors are entirely agglomerated into the same region is stable when the transport cost of the good produced by the modern sector is sufficiently low. Furthermore, we also show that the range of transport costs for which the core–periphery structure is stable expands as the knowledge externalities among skilled workers become more localized.

In both cases, an R&D sector appears to be a strong centripetal force at the multiregional level, thus amplifying the circular causation that lies at the heart of the core–periphery model. Such a result seems to confirm the idea that *growth and agglomeration go hand in hand*. We find it interesting to observe that this result lies at the heart of economic policy debates in several industrialized countries. Indeed, our positive analysis seems to give credit to a trade-off between growth and equity. However, our welfare analysis supports the idea that the additional growth spurred by agglomeration may lead to a Pareto-dominant outcome. Specifically, when the economy moves from dispersion to agglomeration, innovation follows a faster pace. As a consequence, *even those who stay put in the periphery are better off than under dispersion provided that the growth effect triggered by the agglomeration is strong enough*. It should be stressed that this Pareto-optimal property does not require any transfer whatsoever: it is a pure outcome of market interaction.

To be sure, the unskilled workers who live in the core of the economy enjoy a higher level of well-being than those in the periphery. Thus, what appears here is a situation in which everybody can be made better off because agglomeration generates more growth. However, the gap between those who live in the core and in the periphery enlarges. Put differently, the rich get richer as well as the poor, but without ever catching up. Hence, according to Rawls' principle, there can be no conflict between growth and equity because all the unskilled, even those residing in the periphery, can be made better off.

Yet, absolute discrepancies widen across individuals owing to the differential benefits generated within the core region.[4] In other words, agglomeration gives rise to regressive income distribution effects. Such widening welfare gaps may call for corrective policies even though they might in turn hurt growth and, thus, individual welfare. Indeed, the reduction of regional disparities is a major concern in several parts of the world. For example, in the case of the European Union, Article 130a of the Amsterdam Treaty states that the "Community shall aim at reducing disparities between the levels of development of the various regions and the backwardness of the least favoured regions or islands, including rural areas," and thus a clear social cohesion objective exists.

Note, finally, that, in our setting, the regional income discrepancy reflects, at least to a large extent, the spatial distribution of skills. The welfare gap between the core and the periphery arises because of the additional gains generated by the faster growth that the skilled are able to spur by being agglomerated. This in turn makes the unskilled residing in this region better off, even though their productivity is the same as those living in the periphery. The redistributive issue is, therefore, less simple than initial impressions might suggest.

11.2 A MODEL OF AGGLOMERATION AND GROWTH

Our purpose is to analyze a simple setting in which growth is driven by the increase in the number of varieties (which was constant in Chapter 9), whereas the skilled workers who create the blueprints necessary for the production of the new varieties are free to move across regions. Although changes in the population of skilled and unskilled workers could be important, we make the simplifying assumption that each population keeps the same size over time.[5]

11.2.1 The Model

Our model is very similar to the one presented in Section 9.2.5 in which the fixed cost of a firm belonging to the modern sector is expressed in terms of skilled labor, whereas its marginal cost is expressed in terms of unskilled labor. The main difference is an R&D sector in which patents for new varieties are developed. In turn, the fixed cost of a firm producing a given variety is equal to the cost of acquiring the corresponding patent. Because there is no ambiguity, throughout this chapter we will denote time by t, although this symbol has been used so far to describe transport rates.

The economy consists of two regions, A and B, and three production sectors, namely the traditional sector ($\mathbb{T}$), the modern sector ($\mathbb{M}$), and the innovation sector ($\mathbb{R}$). There are two production factors, the low-skilled workers (L) and the high-skilled workers (H). Both the $\mathbb{T}$- and $\mathbb{M}$-sectors use unskilled workers, whereas the $\mathbb{R}$-sector uses skilled workers. Each unskilled worker is endowed with one unit of L-labor per unit of time and is immobile. Every region has the same number of unskilled ($L/2$) over time, where L is a constant. Each skilled worker is endowed with one unit of H-labor and can move between regions at some positive cost (see Section 11.2.3). The total number of skilled workers in the economy is constant over time; without loss of generality, this number is normalized to 1 so that L may be interpreted as the relative size of the unskilled to the skilled. Given this interpretation, it is important to mention that the value of L does not matter for our main results, thus suggesting that our assumption of fixed populations may be less critical than it seems at first sight. Although the total number of skilled is constant over time, we show how growth is made possible through another variable, *the knowledge capital*, which rises together with the number of patents and varieties.

First, we describe consumers' preferences (the location and time arguments are suppressed when no confusion arises from doing so). All workers have the same instantaneous utility function given by

$$u = Q^{\mu} T^{1-\mu} / \mu^{\mu} (1-\mu)^{1-\mu} \qquad 0 < \mu < 1, \tag{11.1}$$

where T is the consumption of the homogeneous $\mathbb{T}$-good, whereas Q stands for the index of the consumption of the $\mathbb{M}$-varieties given by

$$Q = \left[\int_0^M q(i)^{\rho} \, di \right]^{1/\rho} \qquad 0 < \rho < 1.$$

In this expression, M is the total mass of $\mathbb{M}$-varieties available in the global economy at time t, and $q(i)$ represents the consumption of variety $i \in [0, M]$.

The homogeneous $\mathbb{T}$-good is produced under constant returns and perfect competition. Furthermore, this good is costlessly shipped between the two regions, thus enabling us to normalize its price to 1 across time and location. Hence, if ε denotes the expenditure of a consumer at a given time t, and $p(i)$ is the price of variety i, then the consumer's demand functions are as follows:

$$T = (1-\mu)\varepsilon \tag{11.2}$$

$$q(i) = \mu \varepsilon p(i)^{-\sigma} P^{\sigma-1} \qquad i \in [0, M], \tag{11.3}$$

where P is the price index of $\mathbb{M}$-varieties given by

$$P \equiv \left[\int_0^M p(i)^{-(\sigma-1)} di \right]^{-1/(\sigma-1)}. \tag{11.4}$$

Introducing (11.2) and (11.3) into (11.1) yields the indirect utility function

$$v = \varepsilon P^{-\mu}.$$

Thus, compared with the corresponding expressions obtained in Section 9.2.1, we see that the income Y is replaced throughout by the expenditure ε.

Because they are supposed to live indefinitely, consumers need not any longer equalize income and expenditure at each time t.

We now describe the behavior of an arbitrary consumer j in space and time. If this consumer chooses an expenditure path, $\varepsilon_j(t)$ for $t \in [0, \infty)$ such that $\varepsilon_j(t) \geq 0$, and a location path, $r_j(t)$ for $t \in [0, \infty)$ such that $r_j(t) \in \{A, B\}$, then this consumer's indirect utility at time t is given by

$$v_j(t) = \varepsilon_j(t)[P_{r_j(t)}(t)]^{-\mu}, \tag{11.5}$$

where $P_{r_j(t)}(t)$ is the price index of the $\mathbb{M}$-good in region $r_j(t)$ at time t.[6] If $r_j(t_-) \neq r_j(t)$, then he relocates at time t, and we denote by t_h $(h = 1, 2, \ldots)$ the sequence of such moves.[7]

Moving between regions requires various psychological adjustments that negatively affect a migrant. Hence, a consumer who relocates at time t bears a cost $C_m(t)$ expressed in terms of his lifetime utility. Following a standard approach in endogenous growth theory, the lifetime utility of consumer j at time 0 is then defined by

$$U_j(0) = V_j(0) - \sum_h e^{-\gamma t_h} C_m(t_h), \tag{11.6}$$

where $\gamma > 0$ is the subjective discount rate common to all consumers, and

$$V_j(0) \equiv \int_0^{\infty} e^{-\gamma t} \ln[v_j(t)]dt \tag{11.7}$$

is the lifetime utility gross of migration costs (hence, preferences are intertemporally CES with unit elasticity of intertemporal substitution).

The intertemporal allocation of resources is governed by a global and perfectly competitive capital market in which bonds, bearing an interest rate equal to $\nu(t)$ at time t, are traded. The interest rate is common to both regions because the capital market is equally accessible to all consumers and firms, wherever they reside. We must now specify consumer j's intertemporal budget constraint, that is, the present value of expenditure equals wealth. Let $w_{r_j(t)}(t)$ be the wage rate this consumer receives when he resides in $r_j(t)$ at t. Then, the present value of wage income is given by

$$W_j(0) = \int_0^{\infty} e^{-\bar{\nu}(t)t} w_{r_j(t)}(t)\, dt, \tag{11.8}$$

where $\bar{\nu}(t) \equiv (1/t)\int_0^t \nu(\tau)\, d\tau$ is the average interest rate between 0 and t; in (11.8), the term $\exp[-\bar{\nu}(t)]$ converts one unit of income at time t to an equivalent unit at time 0. Using the budget flow constraint, Barro and Sala-i-Martin (1995, 66) have shown that the consumer's intertemporal budget constraint may be written as follows:

$$\int_0^{\infty} \varepsilon_j(t)e^{-\bar{\nu}(t)t} dt = a_j + W_j(0), \tag{11.9}$$

where a_j is the value of the consumer's *initial assets*.

Consider any given location path $r_j(\cdot)$. Then, if $\varepsilon_j(\cdot)$ stands for an expenditure path that maximizes (11.6) subject to (11.9), the first-order condition implies that

$$\dot{\varepsilon}_j(t)/\varepsilon_j(t) = \nu(t) - \gamma \qquad t \geq 0, \tag{11.10}$$

where $\dot{\varepsilon}_j(t) \equiv d\varepsilon_j(t)/dt$. Because (11.10) must hold for every consumer, it is clear that the following relation must hold

$$\dot{E}(t)/E(t) = \nu(t) - \gamma \qquad t \geq 0, \tag{11.11}$$

where $E(t)$ stands for the total expenditure in the economy at time t.

We now turn to the production side of the economy. As noted above, the $\mathbb{T}$-sector operates under constant returns: one unit of $\mathbb{T}$-good is produced using one unit of L-labor. We assume that the expenditure share ($1 - \mu$) on the $\mathbb{T}$-good is sufficiently large for the $\mathbb{T}$-good to be produced always in both regions.[8] In this case, the wage rate of the unskilled is always equal to 1 in each region

$$w_A^L = w_B^L = 1 \qquad t \geq 0, \tag{11.12}$$

for the price of the traditional good is 1 across regions.

In the $\mathbb{M}$-sector, the production of any variety, say i, requires the use of the patent specific to this variety, which has been developed in the $\mathbb{R}$-sector. Once a firm has acquired the patent at the market price (which corresponds to this firm's fixed cost), it can produce one unit of this variety by using one unit of L-labor. The transportation of this variety within the same region is costless. However, when this variety is moved from one region to the other, only a fraction $1/\Upsilon$ arrives at its destination, where $\Upsilon > 1$. Hence, if variety i is produced in region $r = A, B$ and sold at the mill price $p_r(i)$, the price $p_{rs}(i)$ paid by a consumer located in region $s \neq r$ is

$$p_{rs}(i) = p_r(i)\Upsilon. \tag{11.13}$$

Let E_r be the total expenditure in region r at the time in question and P_r be the price index of the $\mathbb{M}$-good in this region. Then, using (11.3) and (11.13), the total demand for variety i produced in region r equals

$$q_r(i) = \mu E_r p_r(i)^{-\sigma} P_r^{\sigma-1} + \mu E_s [p_r(i)\Upsilon]^{-\sigma} P_s^{\sigma-1}\Upsilon, \tag{11.14}$$

where $r, s = A, B$ and $r \neq s$, whereas the last Υ accounts for the melting of the variety during its transportation. The corresponding profit is

$$\pi_r(i) = [p_r(i) - 1]q_r(i),$$

which yields the equilibrium price common to all varieties produced in region r:

$$p_r^* = 1/\rho. \tag{11.15}$$

For notational convenience, we set

$$\phi \equiv \Upsilon^{-(\sigma-1)}.$$

Then, if M_r denotes the number of $\mathbb{M}$-varieties produced in region r at the time in question (which may differ from the number of patents developed in this region), substituting (11.15) into (11.4) yields

$$P_r = (1/\rho)(M_r + M_s\phi)^{-1/(\sigma-1)}, \tag{11.16}$$

where $r, s = A$ or B and $r \neq s$. Furthermore, substituting (11.15) and (11.16) into (11.14), we obtain the equilibrium output of any variety produced in region r:

$$q_r^* = \mu\rho\left(\frac{E_r}{M_r + \phi M_s} + \frac{\phi E_s}{\phi M_r + M_s}\right), \tag{11.17}$$

whereas the equilibrium profit is given by

$$\pi_r^* = q_r^*/(\sigma - 1) \tag{11.18}$$

because

$$\frac{1}{\rho} - 1 = \frac{1}{\sigma - 1}.$$

We now study the labor market clearing conditions for the unskilled. If L_r^M denotes the demand of the unskilled by the $\mathbb{M}$-sector in region r, then

$$L_r^M = M_r q_r^*$$

and, by (11.17),

$$L_A^M + L_B^M = \mu\rho(E_A + E_B),$$

or, setting $E \equiv E_A + E_B$, we obtain

$$L_A^M + L_B^M = \mu\rho E. \tag{11.19}$$

Using (11.2), the total demand for the $\mathbb{T}$-good is $T = (1 - \mu)E$ so that the total demand of L-labor in the $\mathbb{T}$-sector is equal to

$$L^T = (1 - \mu)E. \tag{11.20}$$

In equilibrium, we must have

$$L^T + L_A^M + L_B^M = L,$$

and thus (11.19) and (11.20) imply that, in equilibrium, the total expenditure

$$E^* = \frac{L}{1 - \mu(1 - \rho)} \tag{11.21}$$

is independent of time because L is constant. Therefore, we may conclude from (11.11) that the equilibrium interest rate is equal to the subjective discount rate over time

$$\nu^*(t) = \gamma \quad \text{for all } t \geq 0. \tag{11.22}$$

As a result, from (11.10), the expenditure of any specific consumer j is also a constant, which is readily obtained from (11.9) and (11.22):

$$\varepsilon_j = \gamma[a_j + W_j(0)]. \tag{11.23}$$

11.2.2 The R&D Sector

Turning to the innovation sector, the patents for the new varieties are produced by perfectly competitive laboratories that use skilled workers and benefit from technological spillovers. Following the literature on endogenous growth theory (Romer 1990; Grossman and Helpman 1991, chap. 3), we assume that the productivity of a researcher increases with the total capital of past ideas and methods and that this capital has the nature of a (possibly local) public good. To be specific, when the knowledge capital in region r is K_r, the productivity of each skilled worker residing in r is given by K_r. Recall that the mass of skilled workers in the economy is constant and normalized to 1 ($H_A + H_B = 1$). Hence, when the share of the skilled in region r is λ_r, the number of patents developed per unit of time in region r is such that

$$n_r = K_r \lambda_r. \tag{11.24}$$

Furthermore, it is assumed that the knowledge capital in each region is determined as the outcome of the interactions among *all* skilled workers because each one has something to learn from the others. However, the intensity of these interactions varies with the spatial distribution of skilled workers. More precisely, when worker j has a personal knowledge given by $h(j)$ (e.g., his human capital or the number of papers he has read), the knowledge capital available in region r is given by

$$K_r = \left[\int_0^{\lambda_r} h(j)^\beta dj + \eta \int_0^{1-\lambda_r} h(j)^\beta dj \right]^{1/\beta} \qquad 0 < \beta < 1, \tag{11.25}$$

where β represents an inverse measure of skilled workers' complementarity in knowledge creation, whereas the parameter η ($0 \leq \eta \leq 1$) expresses the intensity of knowledge spillovers between the two regions.

Finally, we assume that worker j's personal knowledge rises with the number of existing patents (e.g., published papers) in the global economy. For simplicity, this knowledge is taken to be proportional to the stock of patents:

$$h(j) = \alpha M.$$

When α is normalized to 1 without loss of generality, it then follows from (11.25) that

$$K_r = M[\lambda_r + \eta(1 - \lambda_r)]^{1/\beta} \qquad 0 < \beta < 1. \tag{11.26}$$

When $\eta = 1$, we have $K_r = M$, which corresponds to the case in which there is no distance-decay effect in knowledge diffusion, and thus knowledge is a pure public good. By contrast, when $\eta = 0$, we have $K_r = Mk(\lambda_r)$, thus implying that knowledge is a local public good. In this way, the parameter η is a measure of the "localness" of knowledge.

As will be seen below, it is not necessary to consider a specific functional form such as (11.26). For our analysis to hold, it is sufficient to assume that

$$K_r = Mk[\lambda_r + \eta(1 - \lambda_r)], \tag{11.27}$$

where $k(\cdot)$ is a strictly convex and increasing function such that

$$k(0) = 0 \quad \text{and} \quad k(1) = 1.$$

Expression (11.27) implies that both regions are in a symmetric relationship in the sense that their own knowledge capital depends only on the distribution of the skilled and not upon their specific attributes. Substituting (11.27) into (11.24) yields

$$n_r = Mk[\lambda_r + \eta(1 - \lambda_r)]\lambda_r. \tag{11.28}$$

The length of patents is assumed to be infinite, and thus a firm that produces a particular variety enjoys a monopoly position forever. This yields the following equation of motion for the number of varieties (or, equivalently, of patents) in the economy:

$$\begin{aligned} \dot{M} &= n_A + n_B \\ &= M\{\lambda k[\lambda + \eta(1 - \lambda)] + (1 - \lambda)\, k(1 - \lambda + \eta\lambda)\}, \end{aligned}$$

where $\lambda \equiv \lambda_A$ and $1 - \lambda \equiv \lambda_B$. For notational simplicity, we set

$$k_A(\lambda) \equiv k[\lambda + \eta(1 - \lambda)] \qquad k_B(\lambda) \equiv k(1 - \lambda + \eta\lambda)$$

and

$$g(\lambda) \equiv \lambda k_A(\lambda) + (1 - \lambda)\, k_B(\lambda). \tag{11.29}$$

Consequently, the preceding equation of motion becomes

$$\dot{M} = g(\lambda)M, \tag{11.30}$$

where $g(\lambda)$ is the *growth rate* of the number of patents and varieties in the global economy when the distribution of skilled workers is λ. It can readily be verified that $g(\lambda)$ is symmetric about $1/2$ and such that

$$g(0) = g(1) = 1,$$

whereas, for $\eta < 1$

$$g'(\lambda) \gtreqless 0 \text{ as } \lambda \gtreqless \frac{1}{2} \quad \text{and } g''(\lambda) > 0 \qquad \lambda \in (0, 1).$$

This implies that, for any given $\eta < 1$, the number of varieties grows at the highest rate when the innovation sector is agglomerated in one region, whereas it grows at the lowest rate when this sector is fully dispersed. For any given function $k(\cdot)$, this rate depends only upon the spatial distribution of skilled workers. It can readily be verified that for $\eta = 1$

$$g(\lambda) = 1 \qquad \lambda \in [0, 1],$$

which corresponds to the normalization of the function $k(\cdot)$ made above, in which case the spatial distribution of the $\mathbb{R}$-sector no longer matters. Furthermore, $g(\lambda)$ is shifted upward when η rises and reaches its maximum value when $\eta = 1$. This means that a distance-decay effect in the diffusion of knowledge slows down the pace of innovation.

We now turn to the formation of wages for the skilled workers. In the production function of patents (11.28), each $\mathbb{R}$-firm located in region r takes the knowledge capital K_r as given. Hence, from such a firm's viewpoint, the marginal productivity of H-labor is equal to K_r. Because the equilibrium wage of a skilled residing in r, denoted by w_r, is given by the average productivity of H-labor in this region, (11.28) implies that the unit costs of a new patent is given by

$$w_r / M k_r(\lambda).$$

Firms enter freely into the $\mathbb{R}$-sector. Hence, if Π_r denotes the market price of a patent developed in region r, the zero-profit condition implies that

$$\Pi_r = w_r / M k_r(\lambda)$$

so that

$$w_r^* = \Pi_r M k_r(\lambda). \tag{11.31}$$

In addition, free entry in the $\mathbb{M}$-sector implies that Π_r equals the asset value of an $\mathbb{M}$-firm that starts producing a new variety by using the corresponding patent. This value, however, cannot be determined without specifying the conditions governing the interregional mobility of patents and, therefore, that of the $\mathbb{M}$-firms. These conditions are discussed in the next two sections.

We may now determine individual expenditure for each type of worker. We assume that all $\mathbb{M}$-firms at time zero are equally shared among the skilled workers.[9] Using (11.23), this implies that $a_L = 0$ and $W_j(0) = \int_0^\infty e^{-\gamma t} dt = 1/\gamma$, and thus

$$\varepsilon_j^* = 1 \qquad j \in L. \tag{11.32}$$

On the other hand, for each skilled, we have

$$\varepsilon_j = \gamma[a_H + W_j(0)] \qquad j \in H, \tag{11.33}$$

where the initial endowment of a skilled is given by

$$a_H = M_A(0)\Pi_A(0) + M_B(0)\Pi_B(0), \tag{11.34}$$

whereas $W_j(0)$ is determined through (11.8) and (11.31) under the specific location path followed by the worker.

11.2.3 Migration Behavior

As in Section 9.5.3, the migration cost is given by

$$C_m(t) = |\dot{\lambda}(t)|/\delta, \tag{11.35}$$

where $\dot{\lambda}(t)$ represents the flow of skilled workers moving from one region to the other and $\delta > 0$ a positive constant. It is positive (negative) when skilled workers move from B to A (from A to B).

Consider the following case that will be relevant for the stability analysis of a steady-state equilibrium at $\tilde{\lambda} \in (0, 1]$. Without loss of generality, let the initial distribution of skilled be lower than $\tilde{\lambda}$. Suppose that $T > 0$ exists such that a flow of skilled from B to A starts at 0 and stops at T. Hence, we have

$$\begin{aligned} &\dot{\lambda}(t) > 0 \qquad t \in (0, T) \\ &\lambda(t) = \tilde{\lambda} \qquad t \geq T. \end{aligned} \tag{11.36}$$

In this case, all the skilled residing in region B are identical except for their migration time. As a result, we can identify them on the basis of their migration time: for each $t \in [0, T)$, denote by $W(0; t)$ the lifetime wage of a skilled worker who migrates from B to A at time t, that is,

$$W(0; t) = \int_0^t e^{-\gamma s} w_B(s) ds + \int_t^\infty e^{-\gamma s} w_A(s) ds. \tag{11.37}$$

Then, using (11.6) and (11.35), we obtain the lifetime utility of such a migrant from

$$U(0; t) = V(0; t) - e^{-\gamma t}\dot{\lambda}(t)/\delta, \tag{11.38}$$

where $V(0; t)$ is the lifetime utility gross of migration costs. Using (11.5) and (11.7), one may determine $V(0; t)$ as follows:

$$\begin{aligned} V(0; t) = &\frac{1}{\gamma}\ln\gamma + \frac{1}{\gamma}\ln[a_H + W(0; t)] \\ &- \mu\left[\int_0^t e^{-\gamma s}\ln[P_B(s)]ds + \int_t^\infty e^{-\gamma s}\ln[P_A(s)]ds\right]. \end{aligned} \tag{11.39}$$

Furthermore, because in equilibrium the skilled residing in region B do not want to delay their migration beyond T (Fukao and Bénabou 1993), it follows that

$$\lim_{t \to T} C_m(t) = 0.$$

Taking the limit of (11.38), we therefore obtain

$$\begin{aligned} U(0;T) &= V(0;T) \\ &= \frac{1}{\gamma}\ln\gamma + \frac{1}{\gamma}\ln[a_H + W(0;T)] \\ &\quad - \mu\left[\int_0^T e^{-\gamma s}\ln[P_B(s)]\,ds + \int_T^\infty e^{-\gamma s}\ln[P_A(s)]\,ds\right]. \end{aligned} \tag{11.40}$$

Because, in equilibrium, all migrants are indifferent about their migration time, it follows that $U(0;t) = U(0;T)$ for all $t \in (0, T)$. Therefore, using (11.38), (11.39), and (11.40), we get

$$\begin{aligned} \dot{\lambda}(t) &= \delta e^{\gamma t}[V(0;t) - V(0;T)] \\ &= \frac{\delta}{\gamma}e^{\gamma t}\ln\left[\frac{a_H + W(0;t)}{a_H + W(0;T)}\right] + \delta\mu e^{\gamma t}\int_t^T e^{-\gamma s}\ln\left[\frac{P_B(s)}{P_A(s)}\right]ds \end{aligned} \tag{11.41}$$

for any $t \in (0, T)$. This expression describes the equilibrium migration dynamics of skilled workers under the expectation (11.36), whereas δ is the speed of adjustment in workers' migration.

In the next sections, the following two polar cases will be considered. In the first, patents are costlessly mobile, presumably because the technology required to produce the new varieties is available everywhere or blueprints developed in a region require no adjustments when used in another region (there is no need for the "tropicalization" of technology). In the second case, patents are perfectly immobile, hence, the corresponding varieties must be produced in the region in which the patents have been developed. This may imply that transferring technologies from one region to the other turns out to be especially hard.

11.3 AGGLOMERATION AND GROWTH WHEN PRODUCTION IS FOOTLOOSE

11.3.1 The Market Outcome

Consider the benchmark case in which firms are footloose in that they are free to produce any new variety anywhere. In other words, a firm producing a specific variety can freely choose its location at each time t regardless of the region where the patent was developed. Therefore, at any given time, if both M_A and M_B are positive, firms' profits at that time must be identical across regions. This in turn implies by (11.18) that $q_A^* = q_B^*$. Hence, using (11.17) as well as $E_A + E_B \equiv E^*$ and $M_A + M_B \equiv M$, we obtain

$$M_A = \frac{E_A - \phi E_B}{(1-\phi)E^*}M \qquad M_B = \frac{E_B - \phi E_A}{(1-\phi)E^*}M, \tag{11.42}$$

and thus

$$M_A > 0 \quad \text{and } M_B > 0 \quad \text{iff } \phi < E_A/E_B < 1/\phi. \tag{11.43}$$

Substituting (11.42) into (11.16) and (11.17), respectively, leads to

$$P_r = (1/\rho)[(1+\phi)(E_r/E^*)M]^{-1/(\sigma-1)} \tag{11.44}$$

and

$$q_A^* = q_B^* = \mu\rho E^*/M. \tag{11.45}$$

In a similar way, it can be shown that

$$M_A = M \quad \text{and} \quad M_B = 0 \qquad \text{iff } E_A/E_B \geq 1/\phi, \tag{11.46}$$

$$P_A = (1/\rho)M^{-1/(\sigma-1)} \qquad P_B = (1/\rho)(\phi M)^{-1/(\sigma-1)}, \tag{11.47}$$

$$q_A^* = \mu\rho E^*/M \geq q_B^* = \mu\rho[\phi E_A + E_B/\phi]/M. \tag{11.48}$$

Likewise, we have

$$M_A = 0 \quad \text{and } M_B = M \qquad \text{iff } E_A/E_B \leq \phi \tag{11.49}$$

$$P_A = (1/\rho)(\phi M)^{-1/(\sigma-1)} \quad \text{and } P_B = (1/\rho)M^{-1/(\sigma-1)} \tag{11.50}$$

$$q_A^* = \mu\rho[E_A/\phi + \phi E_B] \leq q_B^* = \mu\rho E^*/M. \tag{11.51}$$

In the three cases, the equilibrium profit common to all $\mathbb{M}$-firms is given by

$$\pi^* \equiv \max\{\pi_A^*, \pi_B^*\} = \frac{\mu E^*}{\sigma M}, \tag{11.52}$$

in which we have used (11.18), (11.45), (11.48), and (11.51).

11.3.1.1 The ss-Growth Path When λ Is Fixed

To start with, we choose any $\lambda \in [0, 1]$ and study the *steady-state growth path* (in short, the ss-growth path) under that specific λ. Given (11.30), the number of patents (which is equal to the number of $\mathbb{M}$-firms) at time t is such that

$$M(t) = M_0 e^{g(\lambda)t}, \tag{11.53}$$

where M_0 is the initial number of varieties. Using (11.52), we obtain the asset value of a firm at time t as follows:

$$\begin{aligned}\Pi(t) &\equiv \int_t^\infty e^{-\gamma(\tau-t)}\pi^*(\tau)\,d\tau, \\ &= \int_t^\infty e^{-\gamma(\tau-t)}\frac{\mu E^*}{\sigma M(\tau)}\,dt,\end{aligned} \tag{11.54}$$

which is also identical to the equilibrium price of any new patent developed at that time (recall that the place where a patent is developed is immaterial). Hence, the asset value of all firms in the modern sector at time t is such that

$$M(t)\Pi(t) = \frac{\mu E^*}{\sigma} \int_t^\infty e^{-\gamma(\tau - t)} \frac{M(t)}{M(\tau)} d\tau.$$

Because $M(t)/M(\tau) = \exp[-g(\lambda)(\tau - t)]$ by (11.30), we obtain

$$M(t)\Pi(t) = \frac{\mu E^*}{\sigma[\gamma + g(\lambda)]} \equiv a^*(\lambda), \tag{11.55}$$

which is constant over time. Substituting $a^*(\lambda)$ for $M\Pi_r$ in (11.31), therefore, yields the equilibrium wage rate of the skilled in each region:

$$w_A(\lambda) = a^*(\lambda)k[\lambda + \eta(1 - \lambda)] \equiv a^*(\lambda)k_A(\lambda) \tag{11.56}$$

$$w_B(\lambda) = a^*(\lambda)k(1 - \lambda + \eta\lambda) \equiv a^*(\lambda)k_B(\lambda). \tag{11.57}$$

Because, given (11.34), $a_H = a^*(\lambda)$ and $W_j(0) = w_r(\lambda)/\gamma$ for each skilled worker living in region r, (11.33) implies that the total expenditure of all workers in region r at any time equals

$$\begin{aligned} E_r(\lambda) &= \frac{L}{2} + \lambda_r \gamma[a^*(\lambda) + w_r(\lambda)/\gamma] \\ &= \frac{L}{2} + \lambda_r a^*(\lambda)[\gamma + k_r(\lambda)] \end{aligned} \tag{11.58}$$

using (11.56) and (11.57), which leads to

$$\frac{E_A(\lambda)}{E_B(\lambda)} = \frac{L/2 + \lambda a^*(\lambda)[\gamma + k_A(\lambda)]}{L/2 + (1 - \lambda)a^*(\lambda)[\gamma + k_B(\lambda)]}. \tag{11.59}$$

It can then readily be verifiy that

$$\frac{E_A(1)}{E_B(1)} = \frac{\sigma + \mu}{\sigma - \mu} \qquad \frac{E_A(1/2)}{E_B(1/2)} = 1 \qquad \frac{E_A(0)}{E_B(0)} = \frac{\sigma - \mu}{\sigma + \mu}, \tag{11.60}$$

whereas[10]

$$\frac{d[E_A(\lambda)/E_B(\lambda)]}{d\lambda} > 0 \qquad \lambda \in (0, 1). \tag{11.61}$$

Regarding the ss-growth path under the chosen value of λ, there are two different configurations that depend on the value of $\phi \equiv \Upsilon^{-(\sigma-1)}$. First, when the transport cost of the $\mathbb{M}$-good is such that

$$\Upsilon^{\sigma-1} \equiv 1/\phi \geq \frac{\sigma + \mu}{\sigma - \mu}, \tag{11.62}$$

we have the situation depicted in Figure 11.1.

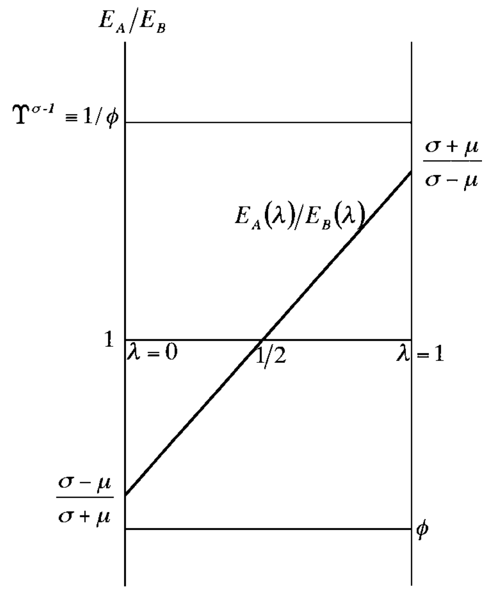


Figure 11.1: The expenditure ratio under high transport costs.

Given (11.60) and (11.61), (11.62) implies that

$$\phi < \frac{E_A(\lambda)}{E_B(\lambda)} < 1/\phi \qquad \lambda \in [0, 1]. \tag{11.63}$$

Hence, it follows from (11.43) that, along the ss-growth path associated with our chosen value of λ, the $\mathbb{M}$-good is always produced in both regions. This should not come as a surprise, for we consider a situation in which the transport cost of this good is high.

Second, when the transport cost of the $\mathbb{M}$-good is such that

$$\Upsilon^{\sigma-1} \equiv 1/\phi \leq \frac{\sigma+\mu}{\sigma-\mu},$$

we have the situation described in Figure 11.2.

We see that, for λ sufficiently close to $1/2$ (that is, when λ is between λ' and λ''), (11.63) holds so that the $\mathbb{M}$-good is produced in both regions. However, when λ is larger than λ'' or smaller than λ', we have

$$\frac{E_A(\lambda)}{E_B(\lambda)} \geq 1/\phi \equiv \Upsilon^{\sigma-1} \quad \text{or} \quad \frac{E_A(\lambda)}{E_B(\lambda)} \leq \phi \equiv \Upsilon^{-(\sigma-1)}.$$

Accordingly, by (11.46) and (11.49), the $\mathbb{M}$-good is entirely produced in the region that has the greater share of the $\mathbb{R}$-sector. Again, this is not very surprising because we are considering the case of low transport cost for the $\mathbb{M}$-good. The

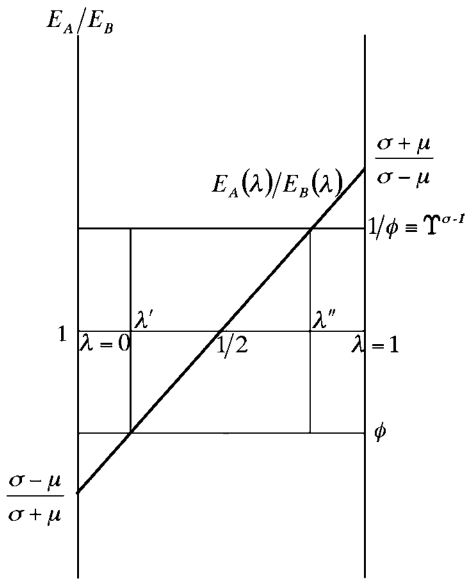


Figure 11.2: The expenditure ratio under low transport costs.

location of the $\mathbb{M}$-sector is then driven by the home-market effect generated by the larger share of skilled workers.

11.3.1.2 The ss-Growth Path When Migration Is Allowed

So far, we have examined the growth path under a fixed distribution of skilled workers between the two regions. Using the results of the preceding sections, we are now equipped to study the steady-state growth path when these workers are free to move but choose not to do so. For that, we must compare the lifetime utility levels of the skilled in the two regions associated with the growth path under any fixed λ and determine the values of λ for which this is an equilibrium.

In (11.6), we will omit the last term because no migration arises in a steady-state equilibrium. Then, for the chosen value of λ, $V_r(0; \lambda)$ stands for the lifetime utility of a skilled worker in region r, whereas $v_r(t; \lambda)$ is the corresponding instantaneous utility at time t. This means that

$$V_r(0; \lambda) = \int_0^\infty e^{-\gamma t} \ln[v_r(t; \lambda)]\, dt, \tag{11.64}$$

so that

$$V_A(0; \lambda) - V_B(0; \lambda) = \int_0^\infty e^{-\gamma t} \ln\left[\frac{v_A(t; \lambda)}{v_B(t; \lambda)}\right] dt. \tag{11.65}$$

From (11.7), (11.33), (11.56), and (11.57), the expenditure of each skilled worker in region r is

$$\varepsilon_r = a^*(\lambda)[\gamma + k_r(\lambda)].$$

Applying (11.5), we get

$$v_r(t;\lambda) = a^*(\lambda)[\gamma + k_r(\lambda)][P_r(t)]^{-\mu}, \tag{11.66}$$

which leads to

$$\frac{v_A(t;\lambda)}{v_B(t;\lambda)} = \frac{\gamma + k_A(\lambda)}{\gamma + k_B(\lambda)}\left[\frac{P_A(t)}{P_B(t)}\right]^{-\mu}. \tag{11.67}$$

Hence, using (11.44), (11.47), and (11.50), respectively, we obtain

$$\left[\frac{P_A(t)}{P_B(t)}\right]^{-\mu} = \left[\frac{E_A(\lambda)}{E_B(\lambda)}\right]^{\mu/(\sigma-1)} \qquad \phi < E_A/E_B < 1/\phi, \tag{11.68}$$

$$\left[\frac{P_A(t)}{P_B(t)}\right]^{-\mu} = \phi^{-\mu/(\sigma-1)} \qquad E_A/E_B \geq 1/\phi, \tag{11.69}$$

$$\left[\frac{P_A(t)}{P_B(t)}\right]^{-\mu} = \phi^{\mu/(\sigma-1)} \qquad E_A/E_B \leq \phi, \tag{11.70}$$

where the expenditure ratio $E_A(\lambda)/E_B(\lambda)$ is given by (11.59). Among other things, these expressions imply that the ratio (11.67) is the same over time when λ is fixed.

Setting

$$\Phi(\lambda) \equiv \frac{v_A(t;\lambda)}{v_B(t;\lambda)},$$

we have

$$V_A(0;\lambda) - V_B(0;\lambda) = \frac{1}{\gamma}\ln\Phi(\lambda) \tag{11.71}$$

and, hence,

$$V_A(0;\lambda) \gtreqless V_B(0;\lambda) \quad \text{as} \quad \Phi(\lambda) \gtreqless 1. \tag{11.72}$$

It can then readily be verified that

$$\Phi(1/2) = 1,$$

and thus

$$V_A(0;1/2) = V_B(0;1/2), \tag{11.73}$$

which implies that full dispersion is always a steady-state equilibrium.

Furthermore, (11.61) means that $E_A(\lambda)/E_B(\lambda)$ increases with λ. Similarly, for $\eta < 1$, $k_A(\lambda)$ is increasing while $k_B(\lambda)$ is decreasing in λ, whereas, for

$\eta = 1$, $k_A(\lambda) = k_B(\lambda) = 1$ for all λ. Thus, for any given $\eta \in [0, 1]$, it follows from (11.67) as well as from (11.68)–(11.70) that

$$\frac{d\Phi(\lambda)}{d\lambda} \geq 0 \qquad \lambda \in (0, 1), \tag{11.74}$$

so that we may conclude by using (11.71) that

$$\frac{d[V_A(0;\lambda) - V_B(0;\lambda)]}{d\lambda} \geq 0 \qquad \lambda \in (0, 1). \tag{11.75}$$

Observe also that, because $\phi < E_A(\lambda)/E_B(\lambda) < 1/\phi$ holds in a neighborhood of $\lambda = 1/2$, (11.74) and (11.75) hold with a strict inequality in that neighborhood. Therefore, we may conclude that

$$V_A(0;\lambda) \gtreqless V_B(0;\lambda) \quad \text{as } \lambda \gtreqless 1/2.$$

These inequalities imply that the economy can be in a steady-state equilibrium under three different values of λ only, that is $\lambda = 1$, $\lambda = 0$, and $\lambda = 1/2$. They also suggest that the equilibrium $\lambda = 1/2$ is unstable, whereas the equilibria $\lambda = 1$ and $\lambda = 0$ are stable. Note, however, that the self-fulfilling nature of the migration process makes stability more difficult to define. Indeed, our model may yield several perfect-foresight solutions under the same initial distribution of skilled workers, λ_0. Consequently, for a given ss-growth path under $\tilde{\lambda}$ (= 0, 1/2, 1), a neighborhood Λ of $\tilde{\lambda}$ may exist such that, for each $\lambda_0 \in \Lambda$, an equilibrium path based on a certain expectation converges to this ss-growth path, whereas another equilibrium path based on another expectation diverges from the same ss-growth path. In this case, is the ss-growth path stable or unstable? A natural way to escape from such a difficulty is to impose a priori some reasonable restriction on the expectations that must be satisfied when an equilibrium path converges to the ss-growth path in question. Because there is perfect foresight, this is equivalent to imposing a restriction on the equilibrium path itself. More precisely, we introduce the following restriction:

Let $\tilde{\lambda} \in [0, 1]$ and $\lambda_0 \in [0, 1]$ such that $\lambda_0 \neq \tilde{\lambda}$. If $\{\lambda(t)\}_{t=0}^{\infty}$ is an equilibrium path satisfying the initial condition $\lambda(0) = \lambda_0$, this path satisfies the *monotonic convergence hypothesis* under $\tilde{\lambda}$ (mc-hypothesis) when $0 < T \leq \infty$ exists such that

$$\text{when } \lambda_0 < \tilde{\lambda} \quad \begin{array}{ll} \dot{\lambda}(t) > 0 & \text{for } t \in (0, T) \\ \lambda(t) = \tilde{\lambda} & \text{for } t \geq T \end{array} \tag{11.76}$$

$$\text{when } \lambda_0 > \tilde{\lambda} \quad \begin{array}{ll} \dot{\lambda}(t) < 0 & \text{for } t \in (0, T) \\ \lambda(t) = \tilde{\lambda} & \text{for } t \geq T. \end{array} \tag{11.77}$$

The ss-growth path under $\tilde{\lambda}$ is said to be *stable* if there exists a neighborhood Λ of $\tilde{\lambda}$ such that, for any $\lambda_0 \in \Lambda$ with $\lambda_0 \neq \tilde{\lambda}$, an equilibrium path exists that satisfies the mc-hypothesis under $\tilde{\lambda}$.[11] The ss-growth path under $\tilde{\lambda}$ is said to be *unstable* when there is no such neighborhood of $\tilde{\lambda}$. Observe that conditions (11.76) and (11.77) are satisfied when the economy moves on a "stable arm" leading to $\tilde{\lambda}$; the same holds when the economy moves on the outer part of a "spiral" leading to $\tilde{\lambda}$ (see Figure 9.6). We show in the appendix that $\tilde{\lambda} = 1/2$ is unstable, whereas $\tilde{\lambda} = 0, 1$ are stable under the mc-hypothesis.

It remains to consider the distribution of the $\mathbb{M}$-sector in the case $\lambda = 1$. The $\mathbb{R}$-sector is entirely agglomerated in region A. However, this is not necessarily true for the location of the $\mathbb{M}$-sector because patents are costlessly mobile. We show below that two different patterns may emerge according to the values of the transport costs of the $\mathbb{M}$-good.

As shown by Figure 11.1, when the transport cost of the $\mathbb{M}$-good is so high that relation (11.62) holds, this good is always produced in the two regions. In particular, using (11.42) and the first relation in (11.60), we see that

$$0 < \frac{M_B(1)}{M_A(1)} = \frac{\sigma - \mu - \phi(\sigma + \mu)}{\sigma + \mu - \phi(\sigma - \mu)} < 1 \quad \text{iff } \Upsilon^{\sigma-1} \equiv 1/\phi > \frac{\sigma + \mu}{\sigma - \mu}. \tag{11.78}$$

We call this spatial configuration a core–periphery pattern of type 1: the core region contains the entire $\mathbb{R}$-sector and the larger share of the $\mathbb{M}$-production sector (but not all of it).

As $\Upsilon^{\sigma-1} \equiv 1/\phi$ decreases toward $(\sigma + \mu)/(\sigma - \mu)$, the ratio M_B/M_A in (11.78) decreases continuously toward zero. Eventually we reach the situation in which

$$M_A = M \quad \text{and} \quad M_B = 0 \quad \text{iff } \Upsilon^{\sigma-1} \equiv 1/\phi \leq \frac{\sigma + \mu}{\sigma - \mu}. \tag{11.79}$$

In this spatial configuration, called a core–periphery pattern of type 2, the core region contains the entire $\mathbb{R}$- and $\mathbb{M}$-sectors.

Consequently, we may conclude as follows:

Proposition 11.1 *When patents are freely mobile, the stable spatial configuration exhibits*

1. *a dominant agglomeration involving the innovation sector entirely and a large fraction of the modern sector in the same region when*

$$\Upsilon^{\sigma-1} > \frac{\sigma + \mu}{\sigma - \mu}; \tag{11.80}$$

2. *a global agglomeration involving the innovation and the modern sectors entirely in the same region when*

$$\Upsilon^{\sigma-1} \leq \frac{\sigma + \mu}{\sigma - \mu} \tag{11.81}$$

As the transport-cost parameter Υ decreases toward 1, the transition from one pattern to the other occurs smoothly.

In either a type-1 or type-2 core–periphery structure, the whole $\mathbb{R}$-sector is agglomerated in the core region. Because the origin of the patents does not matter, R&D firms are able to take full advantage of being agglomerated. Hence, when patents are footloose, the symmetric spatial configuration (in which each region contains half of each $\mathbb{M}$- and $\mathbb{R}$-sector) is never a stable outcome. Such a strong tendency toward concentration is due to the lack of a sufficiently powerful dispersion force. This concept will become clearer after our analysis of the next case.

11.3.2 Should We Mind the Gap?

The foregoing analysis suggests that the pace of growth is faster when agglomeration arises. It is therefore tempting to conclude that there is a conflict between growth and spatial equity because the peripheral region would be a loser when growth is boosted by the agglomeration of mobile activities. This would be so in a zero-sum game, but ours is not. Quite the contrary. As we will see, there might be only gainers in our game, although some regions would gain more than others. This is because global growth may be strong enough for everybody, including the unskilled who live in the peripheral region, to be better off.

To study some of the main aspects of the trade-off between growth and equity, we will assume that the economy is initially on an ss-growth path involving dispersion ($\lambda = 1/2$). From the spatial equity standpoint, this is the best possible outcome because both types of workers reach respectively the same utility level regardless of the region in which they live. Although this outcome is unstable, one could imagine enforcing it by controlling the mobility of the skilled.

Assume now that the economy is left unrestricted, so that any small perturbation will lead it toward a core–periphery structure in which all skilled workers are agglomerated in, say, region A so that $\lambda = 1$. Also assume that, in (11.41), the speed of adjustment (δ) is sufficiently fast for the transition period to be short and, hence, the comparison of the two patterns to be meaningful. There are three groups of individuals to consider: the unskilled residing in regions A and B, respectively, as well as the skilled.

Consider first the case of a core–periphery-structure of type 1 so that transport costs are high in the sense of (11.80). For the unskilled, we know that $w_r^L = \varepsilon_r^L = 1$ for $r = A, B$, and thus (11.5) becomes

$$v_r^L(t;\lambda) = [P_{r_j(t)}(t)]^{-\mu}.$$

Using (11.44), (11.53), (11.55), and (11.58), implies

$$\frac{v_A^L(t;1)}{v_A^L(t;1/2)} = \left(\frac{\sigma+\mu}{\sigma}\right)^{\mu/(\sigma-1)} \exp\left\{\frac{\mu}{\sigma-1}\left[1-k\left(\frac{1+\eta}{2}\right)\right]t\right\},$$

which always exceeds 1 because $\mu > 0$. Hence, using (11.65) for the L-workers

$$V_A^L(0;1) - V_A^L(0;1/2) = \frac{\mu}{\gamma(\sigma-1)}\left[\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} + \ln\left(\frac{\sigma+\mu}{\sigma}\right)\right] > 0, \tag{11.82}$$

and thus the unskilled residing in the core region always prefer agglomeration to dispersion. Regarding the unskilled living in the periphery, we obtain

$$\frac{v_B^L(t;1)}{v_B^L(t;1/2)} = \left(\frac{\sigma-\mu}{\sigma}\right)^{\mu/(\sigma-1)} \exp\left\{\frac{\mu}{\sigma-1}\left[1-k\left(\frac{1+\eta}{2}\right)\right]t\right\}$$

so that

$$V_B^L(0;1) - V_B^L(0;1/2) = \frac{\mu}{\gamma(\sigma-1)}\left[\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} - \ln\left(\frac{\sigma}{\sigma-\mu}\right)\right]. \tag{11.83}$$

The first term inside the bracketed expression stands for the *growth effect* associated with the agglomeration of the $\mathbb{R}$-sector. More precisely, given that $g(1) = k(1) = 1$ and $g(1/2) = k[(1+\eta)/2]$, the numerator of the first term represents the increase in the growth rate of varieties in the economy due to the $\mathbb{R}$-sector agglomeration into the core region; thus, the first term represents the lifetime impact of agglomeration on consumers' welfare. It is strictly positive if and only if $\eta < 1$. The second term represents the disadvantage of being located in the peripheral region, which is measured by the relative increase in the price index of the $\mathbb{M}$-goods in region B. Given (11.83), the unskilled living in the periphery prefer agglomeration to dispersion if and only if

$$\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} > \ln\left(\frac{\sigma}{\sigma-\mu}\right), \tag{11.84}$$

namely, when the extra growth boosted by agglomerating the R&D sector in one region is sufficiently large. This is the more likely, the lower the discount rate (γ), the weaker the spillover effect (η), and the larger the size of the modern sector (μ); on the other hand, more product differentiation (σ falls) enhances the locational disadvantage of the periphery. Thus, the *unskilled residing in the lagging region prefer a core–periphery structure to a dispersed one when the former leads to a sufficiently high rate of growth in the global economy*. In this case, however, there is a welfare gap between the unskilled located in the core and the periphery. Stated differently, growth generates inequalities within the unskilled who are treated differently according to the region in which they live.

Specifically, when $\lambda = 1$, we have

$$\frac{v_A^L(t;1)}{v_B^L(t;1)} = \left(\frac{\sigma+\mu}{\sigma-\mu}\right)^{\mu/(\sigma-1)},$$

and thus the welfare gap is

$$V_A^L(0;1) - V_B^L(0;1) = \frac{\mu}{\gamma(\sigma-1)} \ln\left(\frac{\sigma+\mu}{\sigma-\mu}\right) > 0. \tag{11.85}$$

It remains to consider the skilled workers. Using (11.66), we obtain

$$\frac{v_A^H(t;1)}{v_A^H(t;1/2)} = \frac{v_A^L(t;1)}{v_A^L(t;1/2)}.$$

Thus, when they are agglomerated, the well-being of the skilled increases by the same proportion as the unskilled residing in the core. Indeed,

$$V_A^H(0;1) - V_A^H(0;1/2) = \frac{\mu}{\gamma(\sigma-1)}\left[\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} + \ln\left(\frac{\sigma+\mu}{\sigma}\right)\right] > 0,$$

and thus the skilled always prefer the agglomerated pattern.

Consider now a core–periphery-structure of type 2, thus implying that transport costs are low (see (11.81)). Repeating the same argument as in the foregoing, we obtain the following inequalities:

$$V_A^L(0;1) - V_A^L(0;1/2) = \frac{\mu}{\gamma(\sigma-1)}\left[\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} + \ln\left(\frac{2}{1+\Upsilon^{-(\sigma-1)}}\right)\right],$$

$$V_B^L(0;1) - V_B^L(0;1/2) = \frac{\mu}{\gamma(\sigma-1)}\left[\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} - \ln\left(\frac{1+\Upsilon^{\sigma-1}}{2}\right)\right].$$

In this case, the unskilled living in region *B* prefer the core–periphery structure if and only if

$$\frac{1-k\left(\frac{1+\eta}{2}\right)}{\gamma} > \ln\left(\frac{1+\Upsilon^{\sigma-1}}{2}\right). \tag{11.86}$$

Hence, the lower the transport costs, the more likely the unskilled in the periphery are to be better off under agglomeration.

Finally, under the core–periphery structure, the welfare gap within the unskilled population is given by

$$V_A^L(0;1) - V_B^L(0;1) = \frac{\mu}{\gamma}\ln\Upsilon > 0. \tag{11.87}$$

Summarizing the preceding results, we may conclude as follows.

Proposition 11.2 *Assume that patents are freely mobile. Then, the welfare levels of the three groups of workers (the skilled, the unskilled in region A, and the unskilled in region B) under the core–periphery growth path Pareto-dominate the symmetric growth path if and only if the additional growth boosted by agglomerating the R&D sector in one region is sufficiently large:*

$$\frac{1 - k\left(\frac{1+\eta}{2}\right)}{\gamma} > \ln\left(\frac{\sigma}{\sigma - \mu}\right) \qquad \text{when } \Upsilon^{\sigma-1} > \frac{\sigma + \mu}{\sigma - \mu}$$

or

$$\frac{1 - k\left(\frac{1+\eta}{2}\right)}{\gamma} > \ln\left(\frac{1 + \Upsilon^{\sigma-1}}{2}\right) \qquad \text{when } \Upsilon^{\sigma-1} \leq \frac{\sigma + \mu}{\sigma - \mu}.$$

As expected, at the border case, where

$$\Upsilon^{\sigma-1} = \frac{\sigma + \mu}{\sigma - \mu}$$

(11.84) and (11.86) (as well as (11.85) and (11.87)) are identical.

Finally, note that (11.86) has an interesting implication: when the spillover effect is global ($\eta = 1$), the unskilled residing in region B are always worse off in the core–periphery structure than under dispersion. This is because no growth effect is triggered by agglomeration, whereas the adverse effect of the price index on the unskilled in region B is still there. Consequently, footloose knowledge (if any) would have some unexpected effects: it would negatively affect the extra growth boosted by agglomeration but not the emergence of a core–periphery structure that makes those in the periphery worse off.

11.4 AGGLOMERATION AND GROWTH IN THE PRESENCE OF BARRIERS THAT PREVENT INNOVATION TRANSFER

Let us now focus on the other extreme in which cultural, social, and political barriers prevent the adoption of innovations coming from the other region. In such a society patents developed in one region cannot be transferred to the other; thus, to produce a particular variety of $\mathbb{M}$-good, the corresponding patent must be developed within this region.

Following a similar approach to the one taken in Section 11.3, we first characterize the ss-growth path under any given regional share of skilled workers. Second, we determine which ss-growth path is an equilibrium in which no migration occurs. In the present context, this leads us to compare not only the ratio of utility levels but also the nominal wage rate and price indices in the two regions. Indeed, if one region has a higher nominal wage rate while the other has a lower price index along an ss-growth path, a skilled worker could improve his intertemporal utility by residing in one region for some period of his life and in the other for the rest of the time.[12] We show that an ss-growth path under any asymmetric and interior distribution of skilled workers (i.e., any λ such that $0 < \lambda < 1$ and $\lambda \neq 1/2$) is not migration-proof; thus, only the core–periphery configuration and the symmetric configuration are possible equilibria. Finally, using the migration dynamics, (11.41), we can determine which paths are stable.

11.4.1 The Market Outcome

Under the assumption that patents are nontransferable, the number of $\mathbb{M}$-varieties produced in region r at each time t is equal to the cumulative number of patents previously developed in this region. Thus, using the patent production function (11.28), we get

$$\dot{M}_r = k_r(\lambda)\lambda_r M \qquad r = A, B, \tag{11.88}$$

where $\lambda_A \equiv \lambda$ and $\lambda_B = 1 - \lambda$.

As before, we choose any $\lambda \in [0, 1]$ and study the growth path under that specific λ. In this context, using (11.30), the number of patents available in the economy at time t, given by $M(t) \equiv M_A(t) + M_B(t)$, is still given by (11.53). Hence, rewriting (11.88) leads to

$$\dot{M}_r = k_r(\lambda)\lambda_r M_0 e^{g(\lambda)t} \qquad t \geq 0,$$

where M_0 is the initial number of varieties in the global economy. Solving this differential equation, we obtain

$$M_r(t) = [M_r(0) - \theta_r(\lambda)M_0] + \theta_r(\lambda)M_0 e^{g(\lambda)t}, \tag{11.89}$$

where

$$\theta_r(\lambda) \equiv \frac{k_r(\lambda)\lambda_r}{g(\lambda)} \qquad r = A, B \tag{11.90}$$

represents the share of region r's contribution to the growth in the total number of varieties. Clearly, (11.29) implies that $\theta_A(\lambda) + \theta_B(\lambda) = 1$. It also follows from (11.89) that

$$\lim_{t\to\infty} \frac{M_r(t)}{M(t)} = \theta_r(\lambda) \quad \text{and} \quad \lim_{t\to\infty} \frac{\dot{M}_r(t)}{M_r(t)} = g(\lambda), \tag{11.91}$$

where the convergence process is monotonic.

Using (11.89), we can readily verify that, in each region r, the growth rate is constant over time if and only if

$$M_r(0) = \theta_r(\lambda)M_0. \tag{11.92}$$

In this case, and only in this case, we have

$$\frac{M_r(t)}{M(t)} = \theta_r(\lambda) \quad \text{and} \quad \frac{\dot{M}_r(t)}{M_r(t)} = g(\lambda) \qquad t \geq 0, \tag{11.93}$$

so that

$$M_r(t) = \theta_r(\lambda)M(t) = \theta_r(\lambda)M_0 e^{g(\lambda)t} \qquad t \geq 0. \tag{11.94}$$

In other words, under any fixed $\lambda \in [0, 1]$, an ss-growth path exists if and only if the initial number of patents in each region is given by (11.92). When (11.92) does not hold, (11.91) implies that the growth path under a fixed λ approaches the ss-growth path under λ when $t \to \infty$.

11.4.1.1 ss-Growth Path When λ Is Fixed

To characterize the ss-growth path under any fixed $\lambda \in [0, 1]$, substituting (11.94) into (11.16) and (11.17), while omitting t, yields

$$P_r = (1/\rho)\left\{M[\theta_r(\lambda) + \phi\theta_s(\lambda)]\right\}^{-1/(\sigma-1)}, \tag{11.95}$$

$$q_r^* = \frac{\mu\rho}{M}\left[\frac{E_r}{\theta_r(\lambda) + \phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda) + \theta_s(\lambda)}\right]. \tag{11.96}$$

Consequently, from (11.18) and (11.96), the asset value of an $\mathbb{M}$-firm in region r at time t is given by

$$\begin{aligned}\Pi_r(t) &= \int_t^{\infty} e^{-\gamma(\tau-t)}\pi_r^*(\tau)\,d\tau \\ &= \frac{\mu}{\sigma[\gamma + g(\lambda)]M(t)}\left[\frac{E_r}{\theta_r(\lambda) + \phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda) + \theta_s(\lambda)}\right],\end{aligned} \tag{11.97}$$

where, from (11.32) and (11.33), E_r is given by

$$E_r = \frac{L}{2} + \lambda_r\gamma[a_H + W_r(0)], \tag{11.98}$$

which is independent of time. To determine a_H and $W_r(0)$, we use (11.94) and (11.97) in order to obtain

$$M_r(t)\,\Pi_r(t) = \frac{\mu\theta_r(\lambda)}{\sigma[\gamma + g(\lambda)]}\left[\frac{E_r}{\theta_r(\lambda) + \phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda) + \theta_s(\lambda)}\right], \tag{11.99}$$

which is also independent of time. Substituting (11.99) into (11.34) yields the equilibrium asset value of a skilled $a_H = a^*(\lambda)$ as a function of λ:

$$a^*(\lambda) = \frac{\mu E^*}{\sigma[\gamma + g(\lambda)]}, \tag{11.100}$$

which is identical to (11.55). Furthermore, using (11.31), (11.93), and (11.99) leads to

$$
\begin{aligned}
w_r(t) &= \Pi_r(t)M(t)k_r(\lambda) = \Pi_r(t)M_r(t)k_r(\lambda)/\theta_r(\lambda) \\
&= \frac{\mu k_r(\lambda)}{\sigma[\gamma + g(\lambda)]}\left[\frac{E_r}{\theta_r(\lambda)+\phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda)+\theta_s(\lambda)}\right],
\end{aligned} \tag{11.101}
$$

which is also independent of time. As a result, we have

$$
\begin{aligned}
W_r(0) &= \int_0^{\infty} e^{-\gamma t} w_r(t)\,dt \\
&= \frac{\mu k_r(\lambda)}{\gamma\sigma[\gamma + g(\lambda)]}\left[\frac{E_r}{\theta_r(\lambda)+\phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda)+\theta_s(\lambda)}\right].
\end{aligned} \tag{11.102}
$$

Substituting (11.100) and (11.102) into (11.98) yields

$$
E_r = \frac{L}{2} + \frac{\mu\lambda_r}{\sigma(\gamma+g)}\left\{\gamma E^* + k_r(\lambda)\left[\frac{E_r}{\theta_r(\lambda)+\phi\theta_s(\lambda)} + \frac{\phi E_s}{\phi\theta_r(\lambda)+\theta_s(\lambda)}\right]\right\}
$$

for $r = A, B$. Solving these two linear equations for E_r and E_s, we obtain

$$
E_r^*(\lambda) = \frac{\frac{L}{2} + \frac{\mu g(\lambda)E^*}{\sigma[\gamma+g(\lambda)]}\left[\frac{\gamma}{g(\lambda)}\lambda_r + \frac{\phi\theta_r(\lambda)}{\phi\theta_r(\lambda)+\theta_s(\lambda)}\right]}{1 - \frac{\mu g(\lambda)}{\sigma[\gamma+g(\lambda)]}\frac{(1-\phi^2)\theta_r(\lambda)\theta_s(\lambda)}{[\theta_r(\lambda)+\phi\theta_s(\lambda)][\phi\theta_r(\lambda)+\theta_s(\lambda)]}}, \tag{11.103}
$$

in which $r = A, B, \lambda_A = \lambda$, and $\lambda_B = 1 - \lambda$.

Finally, substituting (11.103) into (11.101) yields the equilibrium wage in region r as a function of λ:

$$
w_r^*(\lambda) = \frac{\mu k_r(\lambda)}{\sigma[\gamma + g(\lambda)]}\left[\frac{E_r^*(\lambda)}{\theta_r(\lambda)+\phi\theta_s(\lambda)} + \frac{\phi E_s^*(\lambda)}{\phi\theta_r(\lambda)+\theta_s(\lambda)}\right], \tag{11.104}
$$

which leads to the equilibrium lifetime wage in region r:

$$
W_r^*(0;\lambda) = \frac{w_r^*(\lambda)}{\gamma}. \tag{11.105}
$$

In turn, using (11.100) and (11.105) in (11.23) yields the equilibrium expenditure of a skilled worker living in region r:

$$
\varepsilon_r^H(\lambda) = \gamma[a^*(\lambda) + W_r^*(0;\lambda)]. \tag{11.106}
$$

11.4.1.2 The ss-Growth Path When Migration Is Allowed

So far, we have examined the ss-growth path under a fixed distribution of skilled workers between the two regions. Using those results, we may determine the equilibrium ss-growth path along which no skilled worker has an incentive to move at any time $t \geq 0$.

For any chosen value of λ, applying (11.5) together with the expenditure function (11.106), we can obtain the indirect utility of each skilled worker in region r at time t:

$$v_r(t;\lambda) = \gamma[a^*(\lambda) + W_r^*(0;\lambda)][P_r(t)]^{-\mu}.$$

Using (11.95), we get

$$\begin{aligned}\Phi(\lambda) \equiv \frac{v_A(t;\lambda)}{v_B(t;\lambda)} &= \frac{a^*(\lambda) + W_A^*(0;\lambda)}{a^*(\lambda) + W_B^*(0;\lambda)}\left[\frac{P_A(t)}{P_B(t)}\right]^{-\mu} \\ &= \frac{a^*(\lambda) + W_A^*(0;\lambda)}{a^*(\lambda) + W_B^*(0;\lambda)}\left[\frac{\theta_A(\lambda) + \phi\theta_B(\lambda)}{\phi\theta_A(\lambda) + \theta_B(\lambda)}\right]^{\mu/(\sigma-1)}, \qquad (11.107)\end{aligned}$$

which is independent of time. Hence, (11.71) and (11.72) still hold.

Therefore, for the ss-growth path under a fixed λ to be migration-proof, it must be that $\Phi(\lambda) = 1$ when $\lambda \in (0, 1)$, whereas $\Phi(\lambda) \geq 1$ when $\lambda = 1$. However, for the reason explained above, this condition is not sufficient. To find a sufficient condition for a migration-proof ss-growth path, we must consider every possible location path of a skilled worker described as follows: let $\varphi(\cdot)$ be a piecewise continuous function on $[0, \infty)$ such that either $\varphi(t) = 1$ or $\varphi(t) = 0$ for each $t \geq 0$, where $\varphi(t) = 1$ means that the skilled worker in question resides in region A at time t, whereas $\varphi(t) = 0$ implies that he resides in region B. Given the ss-growth path under some fixed $\lambda \in [0, 1]$, let $V(\lambda, \varphi(\cdot)) \equiv V(0; \lambda, \varphi(\cdot))$ be the lifetime utility of a skilled worker when he chooses the location path $\varphi(\cdot)$. Using (11.5), (11.7), and (11.23), we then obtain

$$\begin{aligned}V(\lambda, \varphi(\cdot)) &= \frac{1}{\gamma}\ln\gamma + \frac{1}{\gamma}\ln[a^*(\lambda) + W(\lambda, \varphi(\cdot))] \\ &\quad - \mu\int_0^\infty e^{-\gamma t}\left\{\varphi(t)\ln P_A(t) + (1-\varphi(t))\ln P_B(t)\right\}dt \\ &= \frac{1}{\gamma}\ln\gamma + \frac{1}{\gamma}\ln[a^*(\lambda) + W(\lambda, \varphi(\cdot))] \\ &\quad - \mu\left\{\ln\left[p_{A/B}(\lambda)\right]\int_0^\infty e^{-\gamma t}\varphi(t)\,dt + \int_0^\infty e^{-\gamma t}\ln P_B(t)\,dt\right\},\end{aligned}$$

where the lifetime wage income $W(\lambda, \varphi(\cdot))$ and the price index ratio $p_{A/B}(\lambda)$ are defined respectively by

$$\begin{aligned}W(\lambda, \varphi(\cdot)) &\equiv W(0; \lambda, \varphi(\cdot)) \\ &= \int_0^\infty e^{-\gamma t}\varphi(t)\,w_A^*(\lambda)\,dt + \int_0^\infty e^{-\gamma t}[1-\varphi(t)]w_B^*(\lambda)\,dt\end{aligned}$$

and

$$p_{A/B}(\lambda) \equiv \frac{P_A(t)}{P_B(t)} = \left[\frac{\phi\theta_A(\lambda) + \theta_B(\lambda)}{\theta_A(\lambda) + \phi\theta_B(\lambda)}\right]^{1/(\sigma-1)}. \tag{11.108}$$

For convenience, let

$$\bar{\varphi} \equiv \gamma \int_0^\infty e^{-\gamma t}\varphi(t)\,dt$$

be the proportion of the (discounted negative-exponentially) time spent in region A. Denoting $V(\lambda, \bar{\varphi}) \equiv V(\lambda, \varphi(\cdot))$ and $W(\lambda, \bar{\varphi}) \equiv W(\lambda, \varphi(\cdot))$, we may then rewrite these two functions as follows:

$$\begin{aligned} V(\lambda, \bar{\varphi}) \equiv{}& \frac{1}{\gamma}\ln\gamma + \frac{1}{\gamma}\ln[a^*(\lambda) + W(\lambda, \bar{\varphi})] \\ &- \frac{\mu}{\gamma}\bar{\varphi}\ln[p_{A/B}(\lambda)] - \mu\int_0^\infty e^{-\gamma t}\ln P_B(t)\,dt \end{aligned} \tag{11.109}$$

and

$$W(\lambda, \bar{\varphi}) = \frac{1}{\gamma}[\bar{\varphi}w_A^*(\lambda) + (1-\bar{\varphi})w_B^*(\lambda)]. \tag{11.110}$$

Finally, substituting (11.110) into (11.109) yields

$$\begin{aligned} V(\lambda, \bar{\varphi}) ={}& \frac{1}{\gamma}\ln[\gamma a^*(\lambda) + \bar{\varphi}w_A^*(\lambda) + (1-\bar{\varphi})w_B^*(\lambda)] \\ &- \frac{\mu}{\gamma}\bar{\varphi}\ln[p_{A/B}(\lambda)] - \mu\int_0^\infty e^{-\gamma t}\ln P_B(t)\,dt. \end{aligned} \tag{11.111}$$

Hence, for a skilled worker, choosing the optimal location path amounts to choosing the proportion of the time to be spent in region A that maximizes (11.111).

By definition, we have $0 \le \bar{\varphi} \le 1$. Furthermore, $\bar{\varphi} = 1$ if and only if $\varphi(t) = 1$ for all $t \ge 0$; likewise, $\bar{\varphi} = 0$ if and only if $\varphi(t) = 0$ for all $t \ge 0$. Hence, given (11.64), we obtain $V(\lambda, 1) = V_A(0; \lambda)$ and $V(\lambda, 0) = V_B(0; \lambda)$. Therefore, for any given any interior distribution $\lambda \in (0, 1)$, the ss-growth path under λ is migration-proof if and only if

$$V(\lambda, 1) = V(\lambda, 0) = \max\{V(\lambda, \bar{\varphi}); \bar{\varphi} \in [0, 1]\}, \tag{11.112}$$

whereas the ss-growth path under a core–periphery distribution $\lambda = 1$ (say) is migration-proof if and only if

$$V(1, 1) = \max\{V(1, \bar{\varphi}); \bar{\varphi} \in [0, 1]\}. \tag{11.113}$$

In order to examine when condition (11.112) or (11.113) holds, we take the derivatives of (11.111) with respect to $\bar{\varphi}$:

$$\frac{\partial V(\lambda,\bar{\varphi})}{\partial\bar{\varphi}} = \frac{1}{\gamma}\frac{w_A^*(\lambda) - w_B^*(\lambda)}{\gamma a^*(\lambda) + \bar{\varphi}w_A^*(\lambda) + (1-\bar{\varphi})w_B^*(\lambda)} - \frac{\mu}{\gamma}\ln[p_{A/B}(\lambda)] \tag{11.114}$$

$$\frac{\partial^2 V(\lambda,\bar{\varphi})}{\partial\bar{\varphi}^2} = -\frac{1}{\gamma}\frac{[w_A^*(\lambda) - w_B^*(\lambda)]^2}{[\gamma a^*(\lambda) + \bar{\varphi}w_A^*(\lambda) + (1-\bar{\varphi})w_B^*(\lambda)]^2} \leq 0. \tag{11.115}$$

Let us first examine the case of an interior ss-growth path. Given any $\lambda \in (0, 1)$, it must be that $V(\lambda, \cdot)$ is not strictly concave on [0, 1] for (11.112) to hold. Given (11.115), this is so only when

$$w_A^*(\lambda) = w_B^*(\lambda),$$

which implies that $V(\lambda, \cdot)$ is constant on [0, 1]. This, in turn, means that (11.112) holds if and only if (11.114) is zero on [0, 1], thus implying that

$$p_{A/B}(\lambda) = 1, \tag{11.116}$$

namely, the two regions have the same price index. Recalling (11.90) and (11.108), we easily see that (11.116) holds if and only if $\lambda = 1/2$. We may thus conclude as follows:

Proposition 11.3 *Assume that patents cannot be transferred between regions. If a ss-growth path that is not of the core–periphery type is an equilibrium, then it is symmetric.*

In short, when $\lambda \neq 1/2$ and $V(\lambda, 0) = V(\lambda, 1)$, the wage rate is higher in one region while the price index is lower in the other region. In this case, the skilled are able to increase their lifetime utility by "changing places" (formally, they "convexify" their location choices) rather than staying put forever. Such an incentive to change places arises not only from the forward-looking behavior of workers, but also from the presence of saving opportunities, which lead to the averaging of consumption expenditures over time as expressed in (11.23). Note that this result suggests the possibility of an equilibrium growth path along which the cross-migration of skilled workers occurs periodically while the distribution of labor size, λ, is kept constant over time (see Fujita and Thisse 2001 for an elaboration on this point).

Next, we come to the study of the ss-growth path under $\lambda = 1$ (i.e., a core–periphery configuration). Since (11.115) implies that $V(1, \bar{\varphi})$ is concave on [0, 1], (11.113) holds if and only if

$$\left.\frac{\partial V(1,\bar{\varphi})}{\partial\bar{\varphi}}\right|_{\bar{\varphi}=1} \geq 0$$

or, using (11.114), if and only if

$$\frac{w_A^*(1) - w_B^*(1)}{\gamma a^*(1) + w_A^*(1)} \geq \mu \ln[p_{A/B}(1)]. \tag{11.117}$$

Note that

$$\theta_A(1) = 1 \qquad \theta_B(1) = 1 \qquad g(1) = 1 \qquad k_A(1) = 1 \qquad k_B(1) = k(\eta).$$

Hence, using (11.100), (11.103), (11.104), and (11.108), we obtain

$$w_A^*(1) = a^*(1), \tag{11.118}$$

$$w_B^*(1) = a^*(1)\,[k(\eta)/2\sigma]\,[(\sigma+\mu)\phi + (\sigma-\mu)\phi^{-1}], \tag{11.119}$$

$$p_{A/B}(1) = \phi^{1/(\sigma-1)}. \tag{11.120}$$

Substituting (11.118), (11.119), and (11.120) into (11.117) yields

$$\frac{1 - [k(\eta)/2\sigma][(\sigma+\mu)\phi + (\sigma-\mu)\phi^{-1}]}{\gamma+1} \geq \frac{\mu}{\sigma-1}\ln\phi$$

or, equivalently,

$$\Gamma(\phi) \equiv 1 - k(\eta)\left[\frac{1+\mu/\sigma}{2}\phi + \frac{1-\mu/\sigma}{2}\phi^{-1}\right] + \frac{\mu(\gamma+1)}{\sigma-1}\ln\phi^{-1} \geq 0. \tag{11.121}$$

Provided that $\eta > 0$ and, hence, $k(\eta) > 0$,[13] it is readily verified that

$$\Gamma(0) = -\infty \qquad \Gamma(1) \geq 0 \qquad \Gamma'(1) < 0,$$

and that $\Gamma(\phi)$ has a unique inflection point $\phi^* \in (0, 1)$ such that $\Gamma'(\phi^*) = 0$. Therefore, there exists a unique value $\phi_{sustain} \in (0, 1)$ such that

$$\Gamma(\phi_{sustain}) = 0, \tag{11.122}$$

where $\Gamma(\phi) > 0$ for $\phi \in (\phi_{sustain}, 1)$ while $\Gamma(\phi) < 0$ for $\phi < \phi_{sustain}$. Thus, setting

$$\Upsilon_{sustain} \equiv (\phi_{sustain})^{-1/(\sigma-1)},$$

we may conclude as follows:

Proposition 11.4 *When patents cannot be transferred between regions, there always exists a unique sustain point, $\Upsilon_{sustain} > 1$, such that the ss-growth path under a core–periphery structure is an equilibrium if and only if $\Upsilon \leq \Upsilon_{sustain}$.*

It is interesting to compare the nature of the sustain point obtained here with that derived in Section 9.2.5. First, setting $\lambda = 1$ in (11.107) with

$\Phi(1) \equiv \Phi(1; \phi)$ and using (11.118)–(11.120), we have

$$\Phi(1;\phi) = \frac{\gamma a^*(1) + w_A^*(1)}{\gamma a^*(1) + w_B^*(1)} \phi^{-\mu/(\sigma-1)}$$
$$= \frac{\gamma + 1}{\gamma + [k(\eta)/2\sigma][(\sigma+\mu)\phi + (\sigma-\mu)\phi^{-1}]} \phi^{-\mu/(\sigma-1)}. \quad (11.123)$$

As in Chapter 9, suppose that we defined the sustain point, denoted here $\tilde{\phi}$, by the condition

$$\Phi(1;\tilde{\phi}) = 1, \quad (11.124)$$

so that the lifetime utility is the same in the two regions (i.e., $V(1, 1) = V(1, 0)$ when $\phi = \tilde{\phi}$). It then follows from (11.123) that there exists a unique value $\tilde{\phi} \in (0, 1)$ if and only if

$$\mu < \sigma - 1, \quad (11.125)$$

which corresponds to the "no black-hole condition" used in the core-periphery model of Section 9.2.5. However, as seen above, the true sustain point (given by (11.122)) always exists whether (11.125) is met or not. As a consequence, the other parameters being fixed, the core–periphery configuration becomes unsustainable when the transport cost Υ becomes sufficiently large; that is, there is no "black-hole" in the present context.

Second, when condition (11.124) holds, it follows from (11.123) that $w_A^*(1) < w_B^*(1)$, implying that $V(1, \cdot)$ is strictly concave on [0, 1]. Hence, when $\phi = \tilde{\phi}$, (11.112) can never hold so that $\tilde{\phi} < \phi_{sustain}$. Put differently, if we set $\tilde{\Upsilon} \equiv (\tilde{\phi})^{-1/(\sigma-1)}$, we get

$$\Upsilon_{sustain} < \tilde{\Upsilon}.$$

Therefore, because of the incentive to change places, the core-periphery structure is more difficult to sustain in the present context than in the static model of Section 9.2.5, in which no opportunity for savings exists.

Third, because $k(\eta)$ increases with η, it follows from (11.121) and (11.122) that

$$\frac{d\Upsilon_{sustain}}{d\eta} < 0.$$

As a result, when the spillover effect in the $\mathbb{R}$-sector becomes more global, the disadvantage of the periphery in this activity is reduced and, accordingly, the core–periphery structure is sustainable for a smaller range of Υ-values.

Finally, the stability of the ss-growth path under the core–periphery structure can be shown as in the appendix to this chapter whenever $\Upsilon < \Upsilon_{sustain}$. By contrast, the stability analysis of the symmetric ss-growth path is much more involved (see Fujita and Thisse 2001 for more details).

11.4.2 More About the Welfare Gap

The symmetric growth path obtained in this section is identical to that studied in Section 11.3, whereas the ss-growth path corresponding to $\lambda = 1$ is the same as the growth path obtained in the type-2 core–periphery structure (i.e., global agglomeration). Therefore, the welfare implications stated above in the context of the type-2 core–periphery structure are also valid here (see Proposition 11.2). In particular, Pareto-dominance of the ss-growth path under $\lambda = 1$ over the symmetric growth path remains true when (11.86) holds. However, one must keep in mind that here the symmetric growth path may be stable, whereas the core–periphery structure is not always an equilibrium. This implies that market interaction alone does not necessarily lead to the core–periphery growth path when there are barriers to innovation transfers.

11.5 CONCLUDING REMARKS

The results presented in this chapter seem to support Hirschman's and Myrdal's claims quoted in the introduction. When transport costs are sufficiently low, both the modern and the innovation sectors concentrate within the same region, whereas the other region specializes in the production of the traditional good. This is so even though the number of firms operating in the modern sector keeps rising over time regardless of whether technologies are transferable across regions. In fact, our analysis strongly supports the idea that *agglomeration and growth reinforce each other*, confirming and enlarging results obtained in a different context by Martin and Ottaviano (2001). An interesting implication of our analysis is that policies fostering dispersion are likely to hurt global economic growth. Furthermore, the development of footloose technologies makes it even more problematic to prevent the emergence of a core–periphery structure because the symmetric configuration is never a stable equilibrium in this case. By contrast, the existence of barriers to technological transfers may help sustain a dispersed configuration. However, even in this case a deepening of integration is likely to lead to a core–periphery structure.

Nevertheless, the increase of regional disparities does not necessarily imply the impoverishment of the peripheral regions. This would be so when agglomeration does not succeed in boosting enough growth. In this case, the transfer of more economic activities into the core region does hurt those who keep living in the periphery. In the opposite case, it is not so clear that agglomeration, growth, and equity do conflict: people residing in the periphery are better off in the core–periphery structure than under dispersion.[14] There is a conflict only when a fairly narrow interpretation of justice, that is, egalitarianism, is considered because the unskilled living in the core region are better off than those in the periphery. At this stage of the debate, we do not have much to say: the answer depends on societal values. But whatever the answer, it is our contention that understanding regional and urban growth is crucial for improving our knowledge of how modern economies do or may develop.

APPENDIX

When firms are free to produce any variety in any region, we have seen in Section 11.3.1 that the economy can follow a steady-state growth path under three different values of λ, that is, 0, 1/2, and 1. In this appendix, we study the stability of each of these ss-growth paths and show that the ss-growth path under $\tilde{\lambda} = 1/2$ is unstable, whereas it is stable under $\tilde{\lambda} = 0, 1$.

1. To start with, consider the case in which $\tilde{\lambda} = 1/2$. Because the two regions are symmetric, it is sufficient to focus on the values of λ_0 lower than 1/2. The mc-hypothesis then reduces to (11.76), which is itself equivalent to (11.36) in Section 11.2.3. In this case, the equilibrium migration dynamics of skilled workers is given by (11.41). In order to evaluate this expression, we need several preliminary results.

First, recall that the asset value of a firm in the modern sector at time t is given by (11.54). Using (11.30), we obtain

$$M(t)/M(\tau) = e^{-\int_t^\tau g[\lambda(s)]ds}.$$

Hence, at each $t \geq 0$

$$\begin{aligned} a(t) &\equiv M(t)\Pi(t) \\ &= \frac{\mu E^*}{\sigma} \int_t^\infty e^{-\int_t^\tau [\gamma + g(\lambda(s))]\, ds}\, d\tau \end{aligned} \tag{A.1}$$

implying under (11.76) that

$$a(T) = \frac{\mu E^*}{\sigma} \frac{1}{\gamma + g(\tilde{\lambda})}. \tag{A.2}$$

It follows from (A.1) that $a(t)$ is independent of $M(0) = M_0$. As a consequence, M_0 has no influence on the equilibrium values of our variables except $M(t)$.

Next, using (11.31) and setting $\pi_A = \pi_B$, the wage rate of skilled workers in region r at time $t \geq 0$ is given by

$$w_r^*(t) = a(t) k_r[\lambda(t)]. \tag{A.3}$$

Under (11.76), substituting (A.3) into (11.37) yields for $t \leq T$

$$W(0;t) = W(0;T) + \int_t^T e^{-\gamma\tau} a(\tau) \{k_A[\lambda(\tau)] - k_B[\lambda(\tau)]\}\, d\tau, \tag{A.4}$$

where

$$W(0;T)=\int_0^T e^{-\gamma\tau}a(\tau)k_B[\lambda(\tau)]\,d\tau+\frac{a(T)k_A(\tilde{\lambda})}{\gamma}e^{-\gamma T}.$$

By definition, $W(0;t)$ represents the life-time wage of a skilled worker who migrates from B to A at time $t\leq T$. By contrast, the lifetime wage of a skilled worker who stays in region r forever is given by

$$\begin{aligned}W_r(0)&=\int_0^\infty e^{-\gamma t}w_r(t)\,dt\\&=\int_0^\infty e^{-\gamma t}a(t)k_r(\lambda(t))\,dt\qquad r=A,B.\end{aligned}\tag{A.5}$$

Turning now to the aggregate regional expenditure, $E_r(t)$, we use (11.32) and (11.33), and set $a_H=a(0)$. Then, under (11.76), the aggregate expenditure in region A at time $t\leq T$ can be obtained as follows:

$$E_A(t)=\frac{L}{2}+\lambda(0)\gamma\,[a(0)+W_A(0)]+\int_0^t\dot{\lambda}(\tau)\gamma\,[a(0)+W(0;\tau)]\,d\tau,\tag{A.6}$$

where the first two terms represent, respectively, the expenditure of the unskilled and that of the skilled who stay in region A forever, whereas the last term stands for the expenditure by the skilled who have moved from B to A by the time t. Because $E_A(t)+E_B(t)=E^*$, we have

$$E_B(t)=E^*-E_A(t).\tag{A.7}$$

It turns out, however, that another expression of $E_B(t)$ is often more useful. To obtain it, observe that under (11.76), it follows that

$$E_B(T)=\frac{L}{2}+(1-\tilde{\lambda})\gamma[a(0)+W_B(0)],\tag{A.8}$$

whereas differentiating (A.6) and (A.7) with respect to t leads to

$$\dot{E}_B(t)=-\dot{E}_A(t)=-\dot{\lambda}(t)\gamma[a(0)+W(0;t)].$$

Hence, for each $t\leq T$, we get

$$\begin{aligned}E_B(t)&=E_B(T)-\int_t^T\dot{E}_B(\tau)\,d\tau\\&=\frac{L}{2}+(1-\tilde{\lambda})\gamma[a(0)+W_B(0)]+\int_t^T\dot{\lambda}(\tau)\gamma[a(0)+W(0;\tau)]\,d\tau.\end{aligned}\tag{A.9}$$

Putting (A.6) and (A.9) together yields

$$\begin{aligned} E_A(t) - E_B(t) = \lambda_0\gamma[a(0) + W_A(0)] - (1 - \tilde{\lambda})\gamma[a(0) + W_B(0)] \\ + \int_0^t \dot{\lambda}(\tau)\gamma[a(0) + W(0;\tau)]\,d\tau \\ - \int_t^T \dot{\lambda}(\tau)\gamma[a(0) + W(0;\tau)]\,d\tau \qquad t \le T. \end{aligned} \quad \text{(A.10)}$$

We are now ready to establish the following result:

Proposition A.1 *Assume that patents are freely mobile. Then, the ss-growth path under* $\tilde{\lambda} = 1/2$ *is unstable.*

Proof Under (11.76) and $\tilde{\lambda} = 1/2$, we have

$$\lambda(t) < 1/2 \quad \text{for } t < T \quad \lambda(t) = 1/2 \quad \text{for } t \ge T,$$

implying that

$$k_A[\lambda(t)] \equiv k\,\{\lambda(t) + \eta[1 - \lambda(t)]\} \le k\,[1 - \lambda(t) + \eta\lambda(t)] \equiv k_B(\lambda(t)) \quad \text{(A.11)}$$

for $t \ge 0$ because $k(\cdot)$ is increasing and $\eta \le 1$. Furthermore, $a(t) > 0$ for $t \ge 0$ by (A.1). Hence, (A.4) implies that

$$W(0;t) \le W(0;T) \qquad t \le T,$$

whereas $a_H = a(0) > 0$, implying that

$$\frac{a_H + W(0;t)}{a_H + W(0;T)} = \frac{a(0) + W(0;t)}{a(0) + W(0;T)} \le 1 \qquad t \le T. \quad \text{(A.12)}$$

Next, (A.5) and (A.11) together imply that $W_A(0) \le W_B(0)$. Thus, setting $\tilde{\lambda} = 1/2$ in (A.10), we obtain for $t < T$

$$\begin{aligned} E_A(t) - E_B(t) < \left(\lambda_0 - \frac{1}{2}\right)\gamma[a(0) + W_B(0)] \\ + \int_0^T \dot{\lambda}(\tau)\gamma[a(0) + W(0;\tau)]\,d\tau. \end{aligned}$$

Furthermore, for $\tau < T$, it follows from (11.37) and (A.5) that

$$\begin{aligned} W_B(0) - W(0;\tau) &= \int_\tau^\infty e^{-\gamma s}[w_B(s) - w_A(s)]\,ds \\ &= \int_\tau^\infty e^{-\gamma s} a(s)[k_B(\lambda(s)) - k_A(\lambda(s))]\,ds, \end{aligned}$$

which is nonnegative by (A.11). Hence, given that $\lambda(T) = 1/2$, we get

$$E_A(t) - E_B(t) < \left(\lambda_0 - \frac{1}{2}\right)\gamma[a(0) + W_B(0)] + \left(\int_0^T \dot{\lambda}(\tau)d\tau\right)\gamma[a(0) + W_B(0)] = \left(\lambda_0 - \frac{1}{2}\right)\gamma[a(0) + W_B(0)] + \left(\frac{1}{2} - \lambda_0\right)\gamma[a(0) + W_B(0)] = 0$$

or

$$E_A(t) < E_B(t) \qquad t < T.$$

Therefore, using (11.43)–(11.44) and (11.49)–(11.50), we obtain

$$\frac{P_B(t)}{P_A(t)} = \max\left\{\left[\frac{E_A(t)}{E_B(t)}\right]^{1/(\sigma-1)}, \phi^{1/(\sigma-1)}\right\} < 1. \tag{A.13}$$

Inequalities (A.12) and (A.13) imply that the right-hand side of (11.41) is negative for $t < T$, thus contradicting (11.76). Consequently, for any given $\lambda_0 < 1/2$, there is no equilibrium path that satisfies the mc-hypothesis under $\tilde{\lambda} = 1/2$. In other words, that the ss-growth path under $\tilde{\lambda} = 1/2$ is unstable. Q.E.D.

2. Showing the stability of the ss-growth path under $\tilde{\lambda} = 1$ (or $\tilde{\lambda} = 0$) is more involved because we must prove the existence of a neighborhood Λ of $\tilde{\lambda} = 1$ such that, for any $\lambda_0 \in \Lambda$, there is an equilibrium path leading to $\tilde{\lambda} = 1$. We show this through several steps.

First, given $\lambda_0 \in [1/2, 1)$, we assume the existence of an equilibrium path that satisfies (11.76) under $\tilde{\lambda} = 1$ and examine its properties. Observe that under the hypothesis (11.76) and $\tilde{\lambda} = 1$, for any given $\lambda_0 \geq 1/2$ we have

$$1/2 < \lambda(t) < 1 \quad \text{for } t \in (0, T) \qquad \lambda(t) = 1 \quad \text{for } t \geq T, \tag{A.14}$$

implying that

$$k_A[\lambda(t)] \geq k_B[\lambda(t)] \qquad t \geq 0.$$

It then follows from (A.4) that

$$W_A(0) \equiv W(0; 0) \geq W(0; t) \geq W(0; T) \qquad t \leq T, \tag{A.15}$$

which means

$$\frac{a_H + W(0; t)}{a_H + W(0; T)} = \frac{a(0) + W(0; t)}{a(0) + W(0; T)} \geq 1 \qquad t \leq T. \tag{A.16}$$

Furthermore, setting $\tilde{\lambda} = 1$ in (A.10) and using (A.15), we obtain

$$\begin{aligned} E_A(t) - E_B(t) &> \lambda_0 \gamma [a(0) + W_A(0)] - \int_0^T \dot{\lambda}(\tau) \gamma [a(0) + W(0; \tau)] \, d\tau \\ &\geq \lambda_0 \gamma [a(0) + W_A(0)] - \left[\int_0^T \dot{\lambda}(\tau) \, d\tau \right] \gamma [a(0) + W_A(0)] \\ & \qquad t \in (0, T]. \end{aligned}$$

Because

$$\int_0^T \dot{\lambda}(\tau) \, d\tau = 1 - \lambda_0$$

it follows that

$$E_A(t) - E_B(t) > (2\lambda_0 - 1) \gamma [a(0) + W_A(0)] \quad t \in (0, T], \tag{A.17}$$

implying that $E_A(t) > E_B(t)$ when $\lambda_0 \geq 1/2$. Hence, using (11.43)–(11.44) and (11.46)–(11.47) we obtain

$$\frac{P_B(t)}{P_A(t)} = \min \left\{ \left[\frac{E_A(t)}{E_B(t)} \right]^{1/(\sigma - 1)}, (1/\phi)^{1/(\sigma - 1)} \right\} > 1 \qquad t \in (0, T]. \tag{A.18}$$

Substituting each equality in (A.16) and (A.18) into (11.41), let us define for $t \in [0, T]$ that

$$\begin{aligned} \Delta V(t) &\equiv e^{\gamma t} [V(0; t) - V(0; T)] \\ &= \frac{1}{\gamma} e^{\gamma t} \ln \left[\frac{a(0) + W(0; t)}{a(0) + W(0; T)} \right] \\ &\quad + \frac{\mu}{\sigma - 1} e^{\gamma t} \int_t^T e^{-\gamma \tau} \ln \left[\min \left\{ \frac{E_A(\tau)}{E_B(\tau)}, \frac{1}{\phi} \right\} \right] d\tau. \end{aligned} \tag{A.19}$$

Then, given any $\lambda_0 \in [1/2, 1)$, it follows from expressions (A.16), (A.18), and (A.19) that

$$\Delta V(t) > 0 \quad \text{for } t \in [0, T) \quad \text{and } V(T) = 0, \tag{A.20}$$

implying that

$$\dot{\lambda}(t) = \delta \Delta V(t) > 0 \qquad t \in [0, T), \tag{A.21}$$

which is consistent with (11.76). Because $\lambda_0 \in [1/2, 1)$ by assumption and

$$\dot{\lambda}(0) \equiv \lim_{t \to 0} \dot{\lambda}(t) > 0$$

by (A.20), it follows that $\dot{\lambda}(0) > 0$ even when $\lambda_0 = 1/2$, thus demonstrating that expectations do matter.

Next, we show that, starting with any $\lambda_0 \in [1/2, 1)$, the equilibrium path does reach $\tilde{\lambda} = 1$ in a finite time. To do so, first observe by (A.1) and (A.14) that

$$\frac{\mu E^*}{\sigma}\frac{1}{\gamma+1} \leq a(t) \leq \frac{\mu E^*}{\sigma}\frac{1}{\gamma+g(1/2)} \quad \text{for } t \leq T, \tag{A.22}$$

whereas setting $\tilde{\lambda} = 1$ in (A.2) yields

$$a(T) = \frac{\mu E^*}{\sigma}\frac{1}{\gamma+1}. \tag{A.23}$$

In turn, we use (A.5) and (A.14) to obtain

$$W_A(0) \geq \frac{\mu E^*}{\sigma}\frac{1}{\gamma+1}\frac{k\left[\frac{1}{2}(1+\eta)\right]}{\gamma}. \tag{A.24}$$

It also follows from (A.17) that

$$E_B(t) \leq E^*/2 \qquad t \leq T. \tag{A.25}$$

Using (A.22), (A.24), and (A.25) yields

$$\begin{aligned}\frac{E_A(t)}{E_B(t)} &> 1 + \frac{(2\lambda_0 - 1)\,\gamma\,[a(0) + W_A(0)]}{E_B(t)} \\ &\geq 1 + 2(2\lambda_0 - 1)\frac{\mu E^*}{\sigma}\frac{\gamma + k\left[\frac{1}{2}(1+\eta)\right]}{\gamma+1} \qquad t \leq T,\end{aligned}$$

and hence we have by (A.16) and (A.19)

$$\begin{aligned}\Delta V(t) &> \frac{\mu}{\sigma-1}e^{\gamma t}\int_t^T e^{-\gamma s}\ln\left[\min\left\{1 + 2(2\lambda_0 - 1)\right.\right. \\ &\quad \left.\left.\times\frac{\mu E^*}{\sigma}\frac{\gamma + k\left(\frac{1+\eta}{2}\right)}{\gamma+1}, \frac{1}{\phi}\right\}\right] ds \qquad t < T, \\ &= \frac{1 - e^{-\gamma(T-t)}}{\gamma}J(\lambda_0)\end{aligned}$$

where

$$J(\lambda_0) \equiv \frac{\mu}{\sigma-1}\ln\left[\min\left\{1 + 2(2\lambda_0 - 1)\frac{\mu E^*}{\sigma}\frac{\gamma + k\left(\frac{1+\eta}{2}\right)}{\gamma+1}, \frac{1}{\phi}\right\}\right],$$

which is positive for $\lambda_0 > 1/2$ and increasing in λ_0. Hence,

$$\dot{\lambda}(t) = \delta\Delta V(t) > \frac{\delta J(\lambda_0)}{\gamma}\left[1 - e^{-\gamma(T-t)}\right].$$

Integrating both sides from $t = 0$ to T and setting $\lambda(0) = \lambda_0$ and $\lambda(T) = 1$, we get

$$1 - \lambda_0 > \frac{\delta J(\lambda_0)}{\gamma^2}[\gamma T - (1 - e^{-\gamma T})]$$

or

$$\frac{\gamma^2}{\delta}\frac{1 - \lambda_0}{J(\lambda_0)} > \gamma T - (1 - e^{-\gamma T}).$$

Let $T_{\sup}(\lambda_0)$ be the solution to the equation

$$\frac{\gamma^2}{\delta}\frac{1 - \lambda_0}{J(\lambda_0)} = \gamma T - (1 - e^{-\gamma T}). \tag{A.26}$$

Then, it can readily be verified that, for each $\lambda_0 \in (1/2, 1)$, a single solution $T_{\sup}(\lambda_0)$ exists, which is positive, continuous, and decreasing on $(1/2, 1)$, whereas

$$\lim_{\lambda_0 \to 1} T_{\sup}(\lambda_0) = 0.$$

By construction, the value of T associated with the equilibrium path starting with λ_0 is less than $T_{\sup}(\lambda_0)$. Hence, we may conclude as follows:

Lemma A.1 Let $\tilde{\lambda} = 1$ and assume that (11.76) holds. Then, there is a function $T_{\sup}(\lambda_0)$ defined on $(1/2, 1)$, which is positive, continuous, decreasing, and such that the equilibrium path starting with $\lambda_0 \in (1/2, 1)$ at time 0 reaches $\tilde{\lambda} = 1$ before $T_{\sup}(\lambda_0)$, where

$$\lim_{\lambda_0 \to 1} T_{\sup}(\lambda_0) = 0.$$

Because $J(1/2) = 0$, the function $T_{\sup}(\lambda_0)$ defined as the solution to (A.26) has the property

$$\lim_{\lambda_0 \to 1/2} T_{\sup}(\lambda_0) = \infty.$$

However, it can be shown that for $\lambda_0 = 1/2$, the actual time to reach $\tilde{\lambda} = 1$ is finite. This is because $\Delta V(0) > 0$ by (A.20) even when $\lambda_0 = 1/2$, whereas $\Delta V(t)$ is continuous on $[0, T]$. Therefore, along the equilibrium path starting with $\lambda_0 = 1/2$, (A.21) implies that $\lambda(t) > 1/2$ for any small $t > 0$. Then, as in Lemma A.1, we can show that the time required for the path to move from $\lambda(t) > 1/2$ to $\tilde{\lambda} = 1$ is finite, implying that the total time is finite too.

Using the results above, our remaining task is to show the existence of a neighborhood Λ of $\tilde{\lambda} = 1$ such that, for any $\lambda_0 \in \Lambda$, there is an equilibrium path leading to $\tilde{\lambda} = 1$. To do so, it is convenient to express the dynamics of such an equilibrium path by means of differential equations.

Let

$$\varepsilon(t) \equiv \gamma[a(0) + W(0; t)].$$

Then, if

$$(\lambda(t), \Delta V(t), a(t), \varepsilon(t), E_A(t))_{t=0}^{T}$$

is the equilibrium path that starts with the initial distribution λ_0 at time 0 and reaches $\tilde{\lambda} = 1$ at time T, its dynamics can be obtained by using (A.1), (A.4), (A.6), (A.19), and (A.21) as follows: for $t \in (0; T)$

$$\begin{aligned}
\dot{\lambda} &= \delta \Delta V \\
\dot{\Delta V} &= \gamma \Delta V - \frac{a}{\varepsilon}[k_A(\lambda) - k_B(\lambda)] - \frac{\mu}{\sigma - 1} \ln\left[\min\left\{\frac{E_A}{E^* - E_A}, \frac{1}{\phi}\right\}\right] \\
\dot{a} &= [\gamma + g(\lambda)]a - \frac{\mu E^*}{\sigma} \\
\dot{\varepsilon} &= -\gamma e^{-\gamma t} a[k_A(\lambda) - k_B(\lambda)] \\
\dot{E}_A &= \delta \Delta V \varepsilon,
\end{aligned}$$

where the associated terminal conditions can be obtained by using (11.76), (A.7), (A.8), (A.19), and (A.23) as follows:

$$\lambda(0) = \lambda_0 \qquad \lambda(T) = 1 \tag{A.27}$$

$$\begin{aligned}
\Delta V(T) &= 0 \\
a(T) &= \frac{\mu E^*}{\sigma} \frac{1}{1 + \gamma} \\
E_A(T) &= E^* - \frac{L}{2} \\
\varepsilon(T) &= \gamma\, [a(0) + W(0; T)] \\
&= \gamma \left[a(0) + \int_0^T e^{-\gamma \tau} a(\tau) k_B(\lambda(\tau)) d\tau + \frac{\mu E^*}{\gamma \sigma} \frac{e^{-\gamma T}}{\gamma + 1}\right].
\end{aligned} \tag{A.28}$$

The set of terminal conditions derived above is unusual in two respects. First, T is an unknown, whereas λ is specified at both endpoints (see (A.27)). Second, (A.28) is a complex condition involving integrals. Thus, it is not straightforward to show the existence of an equilibrium path starting with each $\lambda_0 \in \Lambda$, where Λ is a neighborhood of $\tilde{\lambda} = 1$. Therefore, we take a slightly different approach to reach the desired result. That is, given that most terminal conditions are specified at $t = T$, we move backward from $t = T$ to $t = 0$ by introducing a new time variable:

$$s \equiv T - t.$$

Furthermore, instead of specifying λ_0, we specify T and then obtain the associated λ_0. That is, using the new variable s, we may rewrite the dynamics as follows: for $s \in (0, T)$ we have

$$\begin{aligned}
\dot{\lambda} &= -\delta \Delta V \\
\Delta\dot{V} &= -\gamma \Delta V + \frac{a}{\varepsilon}\left[k_A(\lambda) - k_B(\lambda)\right] + \frac{\mu}{\sigma - 1} \ln\left[\min\left\{\frac{E_A}{E^* - E_A}, \frac{1}{\phi}\right\}\right] \\
\dot{a} &= -\left[\gamma + g(\lambda)\right] a + \frac{\mu E^*}{\sigma} \qquad \text{(A.29)} \\
\dot{\varepsilon} &= \gamma e^{-\gamma(T-s)} a\left[k_A(\lambda) - k_B(\lambda)\right] \\
\dot{E}_A &= -\delta \Delta V \varepsilon,
\end{aligned}$$

where

$$\begin{aligned}
\lambda(0) &= 1 \\
\Delta V(0) &= 0 \\
a(0) &= \frac{\mu E^*}{\sigma} \frac{1}{1+\gamma} \\
E_A(0) &= E^* - \frac{L}{2} \qquad \text{(A.30)} \\
\varepsilon(0) &= \gamma \left\{a(T) + \int_0^T e^{-\gamma\tau} a(\tau) k_B[\lambda(\tau)]\, d\tau + \frac{\mu E^*}{\gamma\sigma} \frac{e^{-\gamma T}}{\gamma + 1}\right\}. \qquad \text{(A.31)}
\end{aligned}$$

We may then proceed as follows (see Fujita and Thisse 2001 for more details). Because (A.31) is a complex condition, we replace it with

$$\varepsilon(0) = \varepsilon_0, \qquad \text{(A.32)}$$

where ε_0 is a parameter to be chosen appropriately. It can then be shown that, for each $T > 0$ sufficiently small, there exists a closed interval, $I_\varepsilon(T)$, in the positive part of $\mathbb{R}$ such that, for each $\varepsilon_0 \in I_\varepsilon(T)$, the system (A.29), (A.30), and (A.32) has a unique solution denoted by

$$\lambda[(s; T, \varepsilon_0), \Delta V(s; T, \varepsilon_0), a(s; T, \varepsilon_0), \varepsilon(s; T, \varepsilon_0), E_A(s; T, \varepsilon_0)]_{s=0}^{T}.$$

Let $\varepsilon(0; T, \varepsilon_0)$ be the associated value of the right side of (A.31):

$$\varepsilon(0; T, \varepsilon_0) \equiv \gamma \left[a(T; T, \varepsilon_0) + \int_0^T e^{-\gamma\tau} a(\tau; T, \varepsilon_0) k_B(\lambda(\tau; T, \varepsilon_0))\, d\tau + \frac{\mu E^*}{\gamma\sigma} \frac{e^{-\gamma T}}{\gamma + 1}\right].$$

Then, it can be shown that the equation,

$$\varepsilon\,(0; T, \varepsilon_0) = \varepsilon_0$$

has a unique solution, denoted $\varepsilon_0(T)$, which yields the associated value of λ at $s = T$, denoted by

$$\lambda_0(T) \equiv \lambda\,[T; T, \varepsilon_0(T)].$$

Finally, by showing that $\hat{T} > 0$ exists such that $\lambda_0(T)$ is a continuous function on the interval $(0, \hat{T}]$ and

$$\lim_{T \to 0} \lambda_0(T) = 0$$

we obtain the desired neighborhood of $\tilde{\lambda} = 1$, $[\lambda_0(\hat{T}), 1)$. This is sufficient to establish the stability of the ss-growth path under $\tilde{\lambda} = 1$. We may then conclude as follows:

Proposition A.2 *Assume that patents are freely mobile. Then, the ss-growth path under* $\tilde{\lambda} = 1$ *is stable.*

NOTES

1. Around the same time, the same claim was made by Perroux (1955) and Myrdal (1957, chap. 3). For more recent arguments, see Presscott (1998) and Sachs (2000).
2. See Duranton and Puga (2002).
3. Our framework is close to that of Baldwin and Forslid (1997) but is more tractable analytically, thus permitting more concrete results.
4. However, relative discrepancies remain constant.
5. It is worth noting that empirical works cast serious doubt on the idea that growth would be triggered by an increase in the proportion of skilled workers (Jones 1995; Greenwood and Jovanovic 1998).
6. If j is an L-worker, then $r_j(t)$ is either A or B for all t.
7. Thus, an H-worker is allowed to move back and forth several times.
8. A sufficient condition for this to hold is that $1 - \mu > \rho/(1 + \rho)$.
9. Alternatively, we could assume that all the $\mathbb{M}$-firms at time 0 are equally shared by both types of workers. Our results would remain essentially the same.
10. Because $E_A(\lambda) + E_B(\lambda) = E^*$ is constant, it is sufficient to show that $E_A(\lambda)$ increases with λ, which is a property that follows immediately.
11. A neighborhood Λ of $\tilde{\lambda}$ is defined within the subspace $[0, 1]$.
12. In the previous section, this is not an issue because a situation in which one region has a higher nominal wage rate whereas the other has a lower price index along a steady-state growth path never arises.
13. When $\eta = 0$ and, hence, $k(\eta) = 0$, it holds by (11.121) that $F(\phi) > 0$ for any $\phi < 1$, implying that the core-periphery structure is an equilibrium under any $\Upsilon \equiv \phi^{-1} > 1$. That is, when there is no knowledge spillover from the core to the periphery, the R & D activity can never move to the periphery. Hence, the core-periphery structure is sustainable under any $\Upsilon > 1$. The same note applies to Proposition 11.4 in which $\Upsilon_{sustain} = \infty$ when $\eta = 0$. This situation, however, is rather unrealistic; thus, we assume hereafter that $\eta = 0$.
14. In a spatial competition context, Combes and Linnemer (2000) obtain a somewhat similar result: all consumers may be better off under asymmetric equilibrium firms' locations than under symmetric locations. This is because price competition may be fiercer in the former case than in the latter.

Name Index

References

Abdel-Rahman, H. (1996). When do cities specialize in production? *Regional Science and Urban Economics* 26: 1–22.
Abdel-Rahman, H. (2000). City systems: general equilibrium approaches. In: J.-M. Huriot and J.-F. Thisse (eds.). *Economics of Cities. Theoretical Perspectives*. Cambridge: Cambridge University Press, 109–37.
Abdel-Rahman, H., and Fujita, M. (1990). Product variety, Marshallian externalities, and city sizes. *Journal of Regional Science* 30: 165–83.
Acemoglu, D. (1996). A microfoundation for social increasing returns in human capital accumulation. *Quarterly Journal of Economics* 111: 779–804.
Akerlof, G.A. (1997). Social distance and social decisions. *Econometrica* 65, 1005–27.
Alesina A., and Spolaore, E. (1997). On the number and size of nations. *Quarterly Journal of Economics* 112: 1027–55.
Allais, M. (1943). *A la recherche d'une discipline économique*. Reprinted as: *Traité d'économie pure*. Paris: Imprimerie nationale, 1952.
Allen, P.M., and Sanglier, M. (1979). A dynamic model of growth in a central place system. *Geographical Analysis* 11: 256–72.
Allen, P.M., and Sanglier, M. (1981). A dynamic model of central place system – II. *Geographical Analysis* 13: 149–64.
Alonso, W. (1960). Theory of the urban land market. *Papers and Proceedings of the Regional Science Association* 6: 149–57.
Alonso, W. (1964). *Location and Land Use*. Cambridge, MA: Harvard University Press.
Amiti, M. (1998). New trade theory and industrial location in the EU: a survey of evidence. *Oxford Review of Economic Policy* 14: 45–53.
Anas, A. (1983). Discrete choice theory, information theory, and the multinomial logit and gravity models. *Transportation Research B* 17: 13–23.
Anas, A. (1987). *Modeling in Urban and Regional Economics*. Chur: Harwood Academic Publishers.
Anas, A. (1990). Taste heterogeneity and spatial urban structure: the logit model and monocentric theory reconciled. *Journal of Urban Economics* 28: 318–35.
Anas, A., Arnott, R., and Small, K.A. (1998). Urban spatial structure. *Journal of Economic Literature* 36: 1426–64.
Anderson, S.P., and Braid, R.M. (1999). Spatial competition, monopolistic competition,

and optimum product diversity. William Vickrey. *International Journal of Industrial Organization* 17: 953–63.

Anderson, S.P., and de Palma, A. (1988). Spatial price discrimination under heterogeneous products. *Review of Economic Studies* 55: 573–92.

Anderson, S.P., de Palma, A., and Thisse, J.-F. (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Anderson, S.P., and Neven, D. (1991). Cournot competition yields spatial agglomeration. *International Economic Review* 32: 793–808.

Anderson, S.P., Schmitt, N., and Thisse, J.-F. (1995). Who benefits from antidumping legislation? *Journal of International Economics* 38: 321–37.

Arnott, R.J. (1979). Optimal city size in a spatial economy. *Journal of Urban Economics* 6: 65–89.

Arnott, R. (1981). Aggregate land rents and aggregate transport costs. *Economic Journal* 91: 331–47.

Arnott, R. (1995). Time for revisionism on rent control. *Journal of Economic Perspectives* 14 Winter: 99–120.

Arnott, R., and Stiglitz, J. (1979). Aggregate land rent, expenditure on public goods, and optimal city size. *Quarterly Journal of Economics* 93: 471–500.

Arrow, K.J., Chenery, H.B., Minhas, B.S., and Solow, R.M. (1961). Capital-labor substitution and economic efficiency. *Review of Economics and Statistics* 43: 225–50.

Arrow, K., and Debreu, G. (1954). Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–90.

Arthur, W.B. (1990). 'Silicon Valley' locational clusters: when do increasing returns imply monopoly. *Mathematical Social Sciences* 19: 235–51.

Arthur, W.B. (1994). *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: The University of Michigan Press.

Asami, Y. (1990). A determination of bid rents through bidding procedures. *Journal of Urban Economics* 27: 188–211.

Asami, Y., Fujita, M., and Smith, T.E. (1990). On the foundations of land use theory: discrete versus continuous population. *Regional Science and Urban Economics* 20: 473–508.

Asami, Y., Fujita, M., and Thisse, J.-F. (1993). A land capitalization approach to the efficient provision of urban facilities. *Regional Science and Urban Economics* 23: 487–522.

Audretsch, D.B., and Feldman, M.P. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review* 86: 630–40.

Aumann, R.J. (1964). Markets with a continuum of traders. *Econometrica* 32: 39–50.

Bacon, R. (1971). An approach to the theory of consumer shopping behavior. *Urban Studies* 8: 55–64.

Bairoch, P. (1965). La baisse des coûts de transports et le développement économique, *Revue de l'Institut de Sociologie*. 309–32.

Bairoch, P. (1985). *De Jéricho à Mexico. Villes et économie dans l'histoire*. Paris: Gallimard. English translation: *Cities and Economic Development: From the Dawn of History to the Present*. Chicago: The University of Chicago Press (1988).

Bairoch, P. (1993). *Economics and World History: Myths and Paradoxes*. Chicago: The University of Chicago Press.

Baldwin, R.E. (1999). Agglomeration and endogenous capital. *European Economic Review* 43: 253–80.

Baldwin, R.E. (2001). Core-periphery model with forward-looking expectations, *Regional Science and Urban Economics* 31: 21–49.

Baldwin, R.E., and Forslid, R. (2000). The core-periphery model and endogenous growth: stabilising and de-stabilising integration. *Economica* 67: 307–24.

Baldwin, R.E., Martin, Ph., and Ottaviano, G.I.P. (2001). Global income divergence, trade and industrialization: the geography of growth take-offs. *Journal of Economic Growth* 6: 5–37.

Barnett, J. (1986). *The Elusive City. Five Centuries of Design, Ambition and Miscalculation*. New York: Harper and Row.

Barro, R.J., and Sala-i-Martin, X. (1995). *Economic Growth*. New York: McGraw-Hill.

Baumeiter, R.F., and Leary, M.R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin* 117: 497–529.

Baumont, C., and Huriot, J.-M. (2000). Urban economics in retrospect: continuity or change? In: J.-M. Huriot and J.-F. Thisse (eds.). *Economics of Cities. Theoretical Perspectives*. Cambridge: Cambridge University Press, 74–107.

Becattini, G. (1990). The Marshallian industrial district as a socio-economic notion. In: F. Pyke, B. Becattini, and W. Sengenberger (eds.). *Industrial Districts and Inter-firm Cooperation in Italy*. Geneva. International Institute for Labour Studies, 37–51.

Becker, R., and Henderson, J.V. (2000). Intra-industry specialization and urban development. In: J.-M. Huriot and J.-F. Thisse (eds.). *Economics of Cities*. Cambridge: Cambridge University Press, 138–66.

Beckmann, M.J. (1952). A continuous model of transportation. *Econometrica* 20: 643–60.

Beckmann, M.J. (1957). On the distribution of rent and residential density in cities. Yale University, mimeo.

Beckmann, M.J. (1969). On the distribution of urban rent and residential density. *Journal of Economic Theory* 1: 60–8.

Beckmann, M.J. (1972a). Von Thünen revisited: a neoclassical land use model. *Swedish Journal of Economics* 74: 1–7.

Beckmann, M.J. (1972b). Spatial Cournot oligopoly. *Papers and Proceedings of the Regional Science Association* 28: 37–47.

Beckmann, M.J. (1973). Equilibrium models of residential land use. *Regional and Urban Economics* 3: 361–8.

Beckmann, M.J. (1976). Spatial equilibrium in the dispersed city. In: Y.Y. Papageorgiou (ed.). *Mathematical Land Use Theory*. Lexington, MA: Lexington Books, 117–25.

Beckmann, M.J., and Puu, T. (1985). *Spatial economics. Density, potential and flow*. Amsterdam: North-Holland.

Beckmann, M.J., and Thisse, J.-F. (1986). The location of production activities. In: P. Nijkamp (ed.). *Handbook of Regional and Urban Economics, Volume 1*. Amsterdam: North-Holland, 21–95.

Belleflamme, P., Picard, P., and Thisse, J.-F. (2000). An economic theory of regional clusters. *Journal of Urban Economics* 48: 158–84.

Ben-Akiva, M., de Palma, A., and Thisse, J.-F. (1989). Spatial competition with differentiated products. *Regional Science and Urban Economics* 19: 5–19.

Bénabou, R. (1994). Working of a city: location, education and production. *Quarterly Journal of Economics* 106: 619–52.

Bénabou, R. (1996). Equity and efficiency in human capital investment: the local connection. *Review of Economic Studies* 63: 237–64.

Berglas, E. (1976). On the theory of clubs. *Papers and Proceedings of the American Economic Association* 66: 116–21.

Berglas, E., and Pines, D. (1981). Clubs, local public goods and transportation costs. *Journal of Public Economics* 15: 141–62.

Berliant, M. (1985). Equilibrium models with land: a criticism and an alternative. *Regional Science and Urban Economics* 15: 325–40.

Berliant, M., and Fujita, M. (1992). Alonso's discrete population model of land use: efficient allocation and land use. *International Economic Review* 33: 535–66.

Berliant, M., and Konishi, H. (2000). The endogenous formation of a city: population agglomeration and marketplaces in a location-specific production economy. *Regional Science and Urban Economics* 30: 298–324.

Berliant, M., Papageorgiou, Y.Y., and Wang, P. (1990). On welfare theory and urban economics: *Regional Science and Urban Economics* 20: 245–261.

Berliant, M., and Zenou, Y. (2000). Labor differentiation and agglomeration in general equilibrium. Washington University, Department of Economics, mimeo.

Black, D. (2000). Local knowledge spillovers and inequality. University of California at Irvine, Department of Economics, mimeo.

Black, D., and Henderson, V. (1999). A theory of urban growth. *Journal of Political Economy* 107: 252–84.

Blanchard, O., and Katz, L.E. (1992). Regional evolutions. *Brookings Papers on Economic Activity*, 1–61.

Bonanno, G. (1990). General equilibrium theory with imperfect competition. *Journal of Economic Surveys* 4: 297–328.

Borukhov, E., and Hochman, O. (1977). Optimum and market equilibrium in a model of a city without a predetermined center. *Environment and Planning A* 9: 849–56.

Brander, J., and Krugman, P.R. (1983). A 'reciprocal dumping' model of international trade. *Journal of International Economics* 15: 313–21.

Brueckner, J.K., Thisse, J.-F., and Zenou, Y. (1999). Why is Central Paris rich and Downtown Detroit poor? An amenity-based theory. *European Economic Review* 43: 91–107.

Brülhart, M. (1998). Trading places: industrial specialisation in the European Union. *Journal of the Common Market Studies* 36: 319–46.

Buchanan, J.M. (1965). An economic theory of clubs. *Economica* 33: 1–14.

Cain, L.P. (1997). Historical perspective on infrastructure and US economic development. *Regional Science and Urban Economics* 27: 117–138.

Calmette, M.-F., and Le Pottier, J. (1995). Localisation des activités: un modèle bisectoriel avec coûts de transport. *Revue Economique* 46: 900–9.

Cantillon, R. (1755). *Essai sur la nature du commerce en général.* London: Fletcher. English translation by H. Higgs. Reprinted in 1964. New York: A.M. Kelley.

Casetti, E. (1971). Equilibrium land values and population density in an urban setting. *Economic Geography* 47: 16–20.

Casetti, E. (1980). Equilibrium population partitions between urban and agricultural occupations. *Geographical Analysis* 12: 47–54.

Chamberlin, E. (1933). *The Theory of Monopolistic Competition*. Cambridge, MA: Harvard University Press.

Chinitz, B. (1961). Contrasts in agglomeration: New York and Pittsburgh. *Papers and Proceedings of the American Economic Association* 51: 279–89.

Chipman, J.S. (1970). External economies of scale and competitive equilibrium. *Quarterly Journal of Economics* 85: 347–85.

Christaller, W. (1933). *Die Zentralen Orte in Süddeutschland*. Jena: Gustav Fischer Verlag. English translation: *The Central Places of Southern Germany*. Englewood Cliffs, NJ: Prentice-Hall (1966).

Chudak, F.A. (1998). Improved approximation algorithms for uncapacitated facility location. In: R.E. Bixby and R.Z. Rios-Mercado (eds.). *Integer Programming and Combinatorial Optimization*. Berlin: Springer, 180–94.

Ciccone, A., and Hall, R.E. (1996). Productivity and the density of economic activity. *American Economic Review* 86: 54–70.

Coase, R. (1937). The nature of the firm. *Economica* 4: 386–405.

Combes, P.-Ph., and Lafourcade, M. (2001). Transportation cost decline and regional inequalities: evidence from France. CERAS Working paper 01-01, Ecole Nationale des Ponts et Chaussées.

Combes, P.-Ph., and Linnemer, L. (2000). Intermodal competition and regional inequalities. *Regional Science and Urban Economics* 30: 131–184.

Cordella, T., and Grilo, I. (2001). Social dumping and relocation: is there a case for imposing a social clause? *Regional Science and Urban Economics* 31: 643–68.

Cournot, A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. English translation: *Researches into the Mathematical Principles of the Theory of Wealth*. New York: Macmillan (1897).

Cremer, H., de Kerchove, A.-M., and Thisse, J.-F. (1983). An economic theory of regional states and capitals. CORE Discussion Paper N°8323, Université catholique de Louvain.

Cremer H., de Kerchove, A.-M., and Thisse, J.-F. (1985). An economic theory of public facilities in space. *Mathematical Social Sciences* 9: 249–62.

Cronon, W. (1991). *Nature's Megalopolis: Chicago and the Great West*. New York: Norton.

Darnell, A.C. (1990). *The Collected Economics Articles of Harold Hotelling*. Heidelberg: Springer-Verlag.

d'Aspremont C., Gabszewicz, J.J., and Thisse, J.-F. (1979). On Hotelling's "Stability in Competition." *Econometrica* 47: 1045–50.

d'Aspremont C., Gabszewicz, J.J., and Thisse, J.-F. (1983). Product differences and prices. *Economics Letters* 11: 19–23.

d'Aspremont, C., Dos Santos Ferreira, R., and Gérard-Varet, L.-A. (1996). On the Dixit-Stiglitz model of monopolistic competition. *American Economic Review* 86: 623–29.

Debreu, G. (1959). *Theory of Value*. New York: Wiley.

De Fraja, G., and Norman, G. (1993). Product differentiation, pricing policy and equilibrium, *Journal of Regional Science* 33: 343–63.

de la Fuente, A. (1997). The empirics of growth and convergence: a selective survey. *Journal of Economic Dynamics and Control* 21: 23–73.

de la Fuente, A., and Vives, X. (1995). Infrastructure and education as instruments of regional policy: evidence from Spain. *Economic Policy* 20: 13–51.

de Palma, A., Ginsburgh, V., Papageorgiou,Y.Y., and Thisse, J.-F. (1985). The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica* 53: 767–81.

de Palma, A., Liu, Q., and Thisse, J.-F. (1994). Optimal locations on a line with random utilities. *Transportation Science* 28: 63–9.

Desrousseaux, J. (1964). Taille et rendement des industries à l'optimum économique. *Annales de Mines* , 43–7.

Dixit, A.K., and Stiglitz, J.E. (1977). Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Dobkins, L.H., and Ioannides, Y.M. (2000). Dynamic evolution of the size distribution of U.S. cities. In: J.-M. Huriot and J.-F. Thisse (eds.). *Economics of Cities*. Cambridge: Cambridge University Press, 217–60.

Domschke, W., and Drexl, A. (1985). *Location and layout planning: an international bibliography*. Heidelberg: Springer-Verlag.

Dos Santos Ferreira, R., and Thisse, J.-F. (1996). Horizontal and vertical differentiation: the Launhardt model. *International Journal of Industrial Organization* 14: 485–506.

Drèze, J.H. (1974). The pure theory of labour-managed and participatory economies. Part I: certainty. CORE Discussion Paper N°7422, Université catholique de Louvain.

Drèze, J.H. (1985). Labor management and general equilibrium. In: D.C. Jones and J. Svejnar (eds.). *Advances in the Economic Analysis of Participatory and Labor-managed Firms*. London: JAI Press, 3–20.

Drezner, Z., ed. (1995). *Facility Location: A Survey of Applications and Methods*. Heidelberg: Springer-Verlag.

Dunn, E.S. (1954). The equilibrium of land-use pattern in agriculture. *Southern Economic Journal* 21: 173–87.

Duranton, G. (1998). Labor specialization transport costs, and city size. *Journal of Regional Science* 38: 553–73.

Duranton, G., and Puga, D. (2000). Diversity and specialization in cities: why, where and when does it matter? *Urban Studies* 37: 533–55.

Duranton, G., and Puga, D. (2002). Nursery cities: urban diversity, process innovation, and the life-cycle of products. *American Economic Review*, forthcoming.

Eaton, B.C., and Lipsey, R.G. (1977). The introduction of space into the neoclassical model of value theory. In: M. Artis and A. Nobay (eds.). *Studies in Modern Economics*. Oxford: Basil Blackwell, 59–96.

Eaton, B.C., and Lipsey, R.G. (1978). Freedom of entry and the existence of pure profits. *Economic Journal* 88: 455–69.

Eaton, B.C., and Lipsey, R.G. (1979). Comparison shopping and the clustering of homogeneous firms. *Journal of Regional Science* 19: 421–35.

Eaton, B.C., and Lipsey, R.G. (1982). An economic theory of central places. *Economic Journal* 92: 56–72.

Eaton, B.C., and Lipsey, R.G. (1997). *On the Foundations of Monopolistic Competition and Economic Geography*. Cheltenham: Edward Elgar.

Eaton, B.C., and Wooders, M.H. (1985). Sophisticated entry in a model of spatial competition. *Rand Journal of Economics* 16: 282–97.

Eaton, J., and Eckstein, Z. (1997). City and growth. Theory and evidence from France and Japan. *Regional Science and Urban Economics* 27: 443–74.

Eberts, R.W., and McMillen, D.F. (1999). Agglomeration economies and urban public infrastructure. In: P. Cheshire and ZE.S. Milss (eds.). *Handbook of Regional and Urban Economics, Volume 3*. Amsterdam: North-Holland, 1455–95.

Economides, N. (1989). Symmetric equilibrium, existence and optimality in a differentiated product market. *Journal of Economic Theory* 47: 178–94.

The Economist (1995). A survey of cities, July 29.

Ekelund Jr., R.B., and Hébert, R.F. (1999). *Secret Origins of Modern Microeconomics. Dupuit and the Engineers*. Chicago: The University of Chicago Press.

Ellickson, B., Grodal, B., Scotchmer, S., and Zame, W.R. (1999). Clubs and the market. *Econometrica* 67: 1185–1217.

Ellickson, B., and Zame, W. (1994). Foundations for a competitive theory of economic geography. University of California at Los Angeles, Department of Economics, mimeo.

Enke, E. (1951). Equilibrium among spatially separated markets: solution by electric analogue. *Econometrica* 19: 40–7.

Ethier, W. (1982). National and international returns to scale in the modern theory of international trade. *American Economic Review* 72: 389–405.

Faini, R. (1984). Increasing returns, non-traded inputs and regional development. *Economic Journal* 94: 308–23.

Fischer, Cl. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. Chicago: University of Chicago Press.

Feldman, M.P. (1994). *The Geography of Innovation*. Dordrecht: Kluwer Academic Publishers.

Feldman, M.P., and Audretsch, D.B. (1999). Innovation in cities: science-based diversity, specialization and localized competition. *European Economic Review* 43: 409–29.

Feldman, M.P., and Florida, R. (1994). The geographic sources of innovation: technological infrastructure and product innovation in the United States. *Annals of the Association of American Geographers* 84: 210–29.

Flatters, F, Henderson, V., and Mieszkowsi, P. (1974). Public goods, efficiency, and regional fiscal equalization. *Journal of Public Economics* 3: 99–112.

Florian, M., and Los, M. (1982). A new look at static spatial price equilibrium models. *Regional Science and Urban Economics* 12: 579–97.

Francis, R.L., and Mirchandani, P.B., eds. (1990). *Discrete Location Theory*. New York: J. Wiley.

Fujita, M. (1985). Towards general equilibrium models of urban land use. *La Revue Economique* 36: 135–67.

Fujita, M. (1986). A monopolistic competition model of spatial agglomeration: a differentiated product approach. Working Paper N°106, Department of Regional Science, University of Pennsylvania.

Fujita, M. (1988). A monopolistic competition model of spatial agglomeration: a differentiated product approach. *Regional Science and Urban Economics* 18: 87–124.

Fujita, M. (1989). *Urban Economic Theory. Land Use and City Size*. Cambridge: Cambridge University Press.

Fujita, M. (2000). Thünen and the new economic geography. Kyoto Institute of Economic Research, Discussion Paper N°521.

Fujita, M., and Hamaguchi, N. (2001). Intermediate goods and the spatial structure of an economy. *Regional Science and Urban Economics* 31: 79–109.

Fujita, M., and Krugman, P. (1995). When is the economy monocentric? von Thünen and Chamberlin unified. *Regional Science and Urban Economics* 25: 505–28.

Fujita, M., Krugman, P., and Mori, T. (1999). On the evolution of hierarchical urban systems. *European Economic Review* 43: 209–51.

Fujita, M., Krugman, P., and Venables, A.J. (1999). *The Spatial Economy. Cities, Regions and International Trade*. Cambridge, MA: MIT Press.

Fujita, M., and Mori, T. (1995). Structural stability and evolution of urban systems, Working Paper N° 171, Department of Regional Science, University of Pennsylvania.

Fujita, M., and Mori, T. (1997). Structural stability and evolution of urban systems. *Regional Science and Urban Economics* 27: 399–442.

Fujita, M., and Ogawa, H. (1982). Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics* 12: 161–96.

Fujita, M., and Smith, T.E. (1987). Existence of continuous residential land-use equilibria. *Regional Science and urban Economics* 17: 549–94.

Fujita, M., and Smith, T.E. (1990). Additive-interaction models of spatial agglomeration. *Journal of Regional Science* 30: 51–74.

Fujita, M., and Tabuchi, T. (1997). Regional growth in postwar Japan. *Regional Science and Urban Economics* 27: 643–70.

Fujita, M., and Thisse, J.-F. (1986). Spatial competition with a land market: Hotelling and von Thünen unified. *Review of Economic Studies* 53: 819–41.

Fujita, M., and Thisse, J.-F. (1991). Spatial duopoly and residential structure. *Journal of Urban Economics* 30: 27–47.

Fujita, M., and Thisse, J.-F. (1993). Technological linkages and efficient location of indivisible activities: Koopmans-Beckmann and von Thünen unified. *Journal of Urban Economics* 34: 118–41.

Fujita, M., and Thisse, J.-F. (2001). Agglomeration and growth with migration and knowledge externalities, Discussion Paper No. 531, Institute of Economic Research, Kyoto University, Kyoto, Japan.

Fujita, M., Thisse, J.-F., and Zenou, Y. (1997). On the endogeneous formation of secondary employment centers in a city. *Journal of Urban Economics* 41: 337–57.

Fukao, K., and Bénabou, R. (1993). History versus expectations: a comment. *Quarterly Journal of Economics* 108: 535–42.

Gabszewicz, J.J., and Thisse, J.-F. (1986). Spatial competition and the location of firms. In: J.J. Gabszewicz, J.-F. Thisse, M. Fujita and U. Schweizer. *Location Theory*. Chur: Harwood Academic Publishers, 1–71.

Gabszewicz, J.J., and Thisse, J.-F. (1992). Location. In: R.E. Aumann and S. Hart (eds.). *Handbook of Game Theory with Economic Applications, Volume 1*. Amsterdam: North-Holland, 281–304.

Gaspar, J., and Glaeser, E.L. (1998). Information technology and the future of cities, *Journal of Urban Economics* 43, 136–56.

Gehrig, T. (1998). Competing exchanges. *European Economic Review* 42: 277–310.

Giersch, H. (1949). Economic union between nations and the location of industries. *Review of Economic Studies* 17: 87–97.

Ginsburgh, V., Papageorgiou, Y.Y., and Thisse, J.-F. (1985). On existence and stability of spatial equilibria and steady-states. *Regional Science and Urban Economics* 15: 149–58.

Glaeser, E.L. (1998). Are cities dying? *Journal of Economic Perspectives* 12: 139–60.

Glaeser, E.L. Kallal, H.D., Scheinkman, J.A., and Shleifer, A. (1992). Growth in cities. *Journal of Political Economy* 100: 1126–52.

Glaeser, E.L., Kolko, J., and Saiz, A. (2001). Consumer city. *Journal of Economic Geography* 1: 27–50.

Goemans, M.X., and Skutellla, M. (2000). Cooperative facility location games. CORE Discussion Paper N°2000/14, Université catholique de Louvain.

Goldstein, G.S., and Gronberg, T.J. (1984). Economies of scope and economies of agglomeration. *Journal of Urban Economics* 16: 91–104.

Goldstein, G.S., and Moses, L.N. (1975). Interdependence and the location of economic activities. *Journal of Urban Economics* 2: 63–84.

Grandmont, J.-M. (1988). Nonlinear difference equations, bifurcations and chaos: an introduction. CEPREMAP Working paper N°8811.

Greenhut, M.L. (1963). *Microeconomics and the Space-Economy*. Chicago: Scott-Foreman.

Greenhut, M.L. (1981). Spatial pricing in the United States, West Germany and Japan. *Economica* 48: 79–86.

Greenhut, M.L., Norman, G., and Hung, C.-S. (1987). *The Economics of Imperfect Competition. A Spatial Approach*. Cambridge: Cambridge University Press.

Greenwood, J., and Jovanovic, B. (1998). Accounting for growth. NBER Working Paper N°6647.

Grilo I., Shy, O., and Thisse, J.-F. (2001). Price competition when consumer behavior is characterized by conformity or vanity. *Journal of Public Economics* 30: 385–408.

Grimaud, A., and Laffont, J.-J. (1989). Existence of a spatial equilibrium. *Journal of Urban Economics* 25: 213–18.

Grossman, G., and Helpman, E. (1991). *Innovation and Growth in the World Economy*. Cambridge, MA: MIT Press.

Guha, S., and Khuller, S. (1998). Greedy strikes back: Improved facility location algorithms, *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 649–57.

Gupta, B., Pal, D., and Sarkar, J. (1997). Spatial Cournot competition and agglomeration in a model of location choice. *Regional Science and Urban Economics* 27: 261–82.

Haaland, J.I., Kind, H.S., Midelfart Knarvik, K.H., and Torstensson, J. (1999). What determines the geography of Europe? CEPR Discussion Paper N°2072.

Hägerstrand, T. (1953). *Innovation Diffusion as a Spatial Process*. Chicago: The University of Chicago Press.

Hall, R.E., and Jones, C.I. (1999). Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116.

Hamilton, B.W. (1980). Indivisibilities and interplant transportation cost: do they cause market breakdown? *Journal of Urban Economics* 7: 31–41.

Hamilton, J., Thisse, J.-F., and Zenou, Y. (2000). Wage competition with heterogeneous workers and firms. *Journal of Labor Economics* 18: 453–72.

Hansen, N. (1990). Do producer services induce regional development? *Journal of Regional Science* 30: 465–78.

Hansen, P., Labbé, M., Peeters, D., and Thisse, J.-F. (1987). Facility location analysis. In: P. Hansen, M. Labbé, D. Peeters, J.-F. Thisse, and J.V. Henderson. *Systems of Cities and Facility Location*. Chur: Harwood Academic Publishers, 1–70.

Hanson, G.H. (1996). Localization economies, vertical organization, and trade. *American Economic Review* 86: 1266–78.

Harris, C. (1954). The market as a factor on the localization of industry in the United States. *Annals of the Association of American Geographers* 64: 315–48.

Hart, O. (1985). Imperfect competition in general equilibrium: an overview of recent work. In K. Arrow and S. Honkapohja (eds.). *Frontiers in Economics*. Oxford: Basil Blackwell, 100–49.

Hartwick, J.M., Schweizer, U., and Varaiya, P.V. (1976). Comparative statics of a residential economy with several classes. *Journal of Economic Theory* 13: 396–413.

Hayek, F.A. (1988). *The Fatal Conceit. The Errors of Socialism*. London: Routledge.

Head, K., and Mayer, T. (2000). Non-Europe. The magnitude and causes of market fragmentation in the EU. *Weltwirtschaftliches Archiv* 136: 284–314.

Heffley, D.R. (1972). The quadratic assignment problem: a note. *Econometrica* 40: 1155–63.

Heffley, D.R. (1976). Efficient spatial allocation in the quadratic assignment problem. *Journal of Urban Economics* 3: 309–22.

Helpman, E. (1998). The size of regions. In: D. Pines, E. Sadka, and I. Zilcha (eds.). *Topics in Public Economics. Theoretical and Applied Analysis.* Cambridge: Cambridge University Press, 33–54.

Helsley, R.W., and Strange, W.C. (1990). Matching and agglomeration economies in a system of cities. *Regional Science and Urban Economics* 20: 189–212.

Helsley, R.W., and Strange, W.C. (1997). Limited developers. *Canadian Journal of Economics* 30: 329–48.

Henderson, J.V. (1972). Hierarchy models of city size: an economic evaluation. *Journal of Regional Science* 12: 435–41.

Henderson, J.V. (1974). The sizes and types of cities. *American Economic Review* 64: 640–56.

Henderson, J.V. (1977). *Economic Theory and the Cities*. New York: Academic Press.

Henderson, J.V. (1985). The Tiebout model: bring back the entrepreneurs. *Journal of Political Economy* 93: 248–64.

Henderson, J.V. (1987). Systems of cities and inter-city trade. In: P. Hansen, M. Labbé, D. Peeters, J.-F. Thisse, and J.V. Henderson. *Systems of Cities and Facility Location*. Chur: Harwood Academic Publishers, 71–119.

Henderson, J.V. (1988). *Urban Development. Theory, Fact and Illusion*. Oxford: Oxford University Press.

Henderson, J.V. (1997a). Medium size cities. *Regional Science and Urban Economics* 27: 583–612.

Henderson, J.V. (1997b). Externalities and industrial development. *Journal of Urban Economics* 42: 449–70.

Henderson, V., Kuncoro, A., and Turner, M. (1995). Industrial development in cities. *Journal of Political Economy* 103: 1066–90.

Henderson, J.V., and Mitra, A. (1996). The new urban landscape: developers and edge cities. *Regional Science and Urban Economics* 26: 613–43.

Henderson, J.V., and Thisse, J.-F. (2001). On strategic community development. *Journal of Political Economy* 109: 546–69.

Herbert, J.D., and Stevens, B.H. (1970). A model of the distribution of residential activity in urban areas. *Journal of Regional Science* 2: 21–36.

Hildenbrand, W. (1974). *Core and Equilibria in a Large Economy*. Princeton, NJ: Princeton University Press.

Hirschman, A.O. (1958). *The Strategy of Development*. New Haven, CN: Yale University Press.

Hochman, O. (1981). Land rents, optimal taxation and local fiscal independence in an economy with local public goods. *Journal of Public Economics* 15: 59–85.

Hochman, O., Pines, D., and Thisse, J.-F. (1995). On the optimal structure of local governments. *American Economic Review* 85: 1224–40.

Hohenberg, P., and Lees, L.H. (1985). *The Making of Urban Europe (1000–1950)*. Cambridge, MA: Harvard University Press.

Hoover, E.M. (1936). *Location Theory and the Shoe and Leather Industries*. Cambridge, MA: Harvard University Press.

Hoover, E.M. (1937). Spatial price discrimination. *Review of Economic Studies* 4: 182–91.

Hoover, E.M. (1948). *The Location of Economic Activity*. New York: McGraw-Hill.

Horn, H., and Shy, O. (1996). Bundling and international market segmentation, *International Economic Review* 37: 51–69.

Hotelling, H. (1929). Stability in competition. *Economic Journal* 39: 41–57.

Hotelling, H. (1938). The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica* 6: 242–69.

Hsu, S.K. (1979). Monopoly output under alternative spatial pricing techniques: comment. *American Economic Review* 69: 678–9.

Huriot, J.-M. (1981). Rente foncière et modéle de production. *Environment and Planning A* 13: 1125–49.

Huriot, J.-M. (1994). *Von Thünen: économie et espace*. Paris: Economica.

Huriot, J.-M., Smith, T., and Thisse, J.-F. (1989). Minimum-cost distance in spatial analysis. *Geographical Analysis* 21: 294–315.

Huriot, J.-M., and Thisse, J.-F., eds. (2000). *Economics of Cities. Theoretical Perspectives*. Cambridge: Cambridge University Press.

Imai, H. (1982). CBD hypothesis and economies of agglomeration. *Journal of Economic Theory* 28: 275–99.

Irmen, A., and Thisse, J.-F. (1998) Competition in multi-characteristics spaces: Hotelling was almost right, *Journal of Economic Theory* 78: 76–102.

Isard, W. (1949). The general theory of location and space-economy. *Quarterly Journal of Economics* 63: 476–506.

Isard, W. (1954). Location theory and trade theory: short-run analysis. *Quarterly Journal of Economics* 68: 305–20.

Isard, W. (1956). *Location and Space-Economy*. Cambridge, MA: MIT Press.

Isard, W., and Peck, M.J. (1954). Location theory and international and interregional trade theory. *Quarterly Journal of Economics* 68: 97–114.

Jacobs, J. (1969). *The Economy of Cities*. New York: Random House.

Jaffe, A.B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–98.

Jones, C.I. (1995). Times series tests of endogenous growth models, *Quarterly Journal of Economics* 110: 495–525.

Justman, M. (1994). The effects of local demand on industry location. *Review of Economics and Statistics* 76: 742–53.

Kaldor, N. (1935). Market imperfection and excess capacity. *Economica* 2: 35–50.

Kaldor, N. (1970). The case for regional policies. *Scottish Journal of Political Economy* 17: 337–48.

Kaldor, N. (1985). *Economics without Equilibrium*. Armonk, NY: M.E. Share.

Kanemoto Y. (1980). *Theory of Urban Externalities.* Amsterdam: North-Holland.

Kanemoto, Y. (1990). Optimal cities with indivisibilities in production and interactions between firms. *Journal of Urban Economics* 27: 46–59.

Kats, A. (1995). More on Hotelling's Stability in Competition. *International Journal of Industrial Organization* 13: 89–93.

Kilkenny, M. (1998). Transport costs and rural development. *Journal of Regional Science* 38: 293–312.

Kolko, J. (1999). Can I get some service here? Information technology, service industries, and the future of cities. Harvard University, Department of Economics, mimeo.

Koopmans, T.C. (1957). *Three Essays on the State of Economic Science*. New York: McGraw-Hill.

Koopmans, T.C., and Beckmann, M.J. (1957). Assignment problems and the location of economic activities. *Econometrica* 25: 1401–14.

Krugman, P. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9: 469–79.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade, *American Economic Review* 70: 950–9.

Krugman, P. (1991a). Increasing returns and economic geography. *Journal of Political Economy* 99: 483–99.

Krugman, P. (1991b). *Geography and Trade*. Cambridge, MA: MIT Press.

Krugman, P. (1991c). History versus expectations. *Quarterly Journal of Economics* 106: 651–67.

Krugman, P. (1995). *Development, Geography, and Economic Theory*. Cambridge MA: MIT Press.

Krugman, P. (1998). Space: the final frontier. *Journal of Economic Perspectives* 12: 161–74.

Krugman, P., and Venables, A.J. (1995). Globalization and the inequality of nations. *Quarterly Journal of Economics* 110: 857–80.

Kuehn, A.A., and Hambuger, M.J. (1963). A heuristic program for locating warehouses. *Management Science* 9: 643–66.

Labbé M., Peeters, D., and Thisse, J.-F. (1995). Location on networks. In: M. Ball, T. Magnanti, C. Monma, and G. Nemhauser (eds.). *Handbook of Operations Research and Management Science: Networks*. Amsterdam: North-Holland, 551–624.

Labys, W.C., and Yang, C.W. (1997). Spatial price equilibrium as a foundation to unified spatial commodity modeling. *Papers in Regional Science* 76: 199–228.

Lancaster, K. (1979). *Variety, Equity and Efficiency*. Oxford: Basil Blackwell.

Launhardt, W. (1885). *Mathematische Begründung der Volkswirtschaftslehre*. Leipzig: B.G. Teubner. English translation: *Mathematical Principles of Economics*. Aldershot: Edward Elgar (1993).

Lea, A.C. (1979). Welfare theory, public goods, and public facility location. *Geographical Analysis* 11: 217–39.

Lederer, P.J., and Hurter, A.P. (1986). Competition of firms: discriminatory pricing and location. *Econometrica* 54: 623–40.

Lepetit, B. (1988). *Les villes dans la France moderne (1740–1840)*. Paris: Albin Michel.

Lerner, A., and Singer, H.W. (1937). Some notes on duopoly and spatial competition. *Journal of Political Economy* 45: 145–86.

Liu, H.-L., and Fujita, M. (1991). A monopolistic competition model of spatial agglomeration with variable density. *Annals of Regional Science* 25: 81–99.

Lösch, A. (1940). *Die Räumliche Ordnung der Wirtschaft*. Jena: Gustav Fischer. English translation: *The Economics of Location*. New Haven, CN: Yale University Press (1954).

Lucas, R.E. (1988). On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–22.

Lucas, R.E. (1990). Why doesn't capital flow from rich to poor countries? *Papers of the American Economic Association* 80: 92–6.

Lucas, R.E. (2001). Externalities and cities. *Review of Economic Dynamics* 4: 245–74.

Ludema, R.D., and Wooton, I. (2000). Economic geography and the fiscal effects of regional integration. *Journal of International Economics* 52: 331–52.

Malinvaud, E. (1968). *Leçons de théorie microéconomique*. Paris: Dunod. English translation: *Lessons in Microeconomic Theory*. Amsterdam: North Holland (1970).

Manne, A. (1964). Plant location under economies of scale – Decentralization and computation. *Management Science* 11: 213–235.

Marchand, B. (1993). *Paris, histoire d'une ville*. Paris: Editions du Seuil.

Markusen, J.R. (1983). Factor movements and commodity trade as complements. *Journal of International Economics* 14: 341–56.

Marshall, A. (1890). *Principles of Economics*. London: Macmillan. 8th edition published in 1920.

Marshall, J.U. (1989). *The Structure of Urban Systems*. Toronto: University of Toronto Press.

Martin, Ph., and Ottaviano, G.I.P. (1999). Growing locations: industry location in a model of endogenous growth. *European Economic Review* 43: 281–302.

Martin, Ph., and Ottaviano, G.I.P. (2001). Growth and agglomeration. *International Economic Review*, forthcoming.

Martin, R. (1999). Editorial: The 'new' economic geography: challenge or irrelevance. *Transactions of the Institute of British Geographers* 24: 389–91.

Martinez-Giralt, X., and Neven, D. (1988). Can price competition dominate market segmentation? *Journal of Industrial Economics* 36: 431–42.

Matsuyama, K. (1995). Complementarities and cumulative process in models of monopolistic competition. *Journal of Economic Literature* 33: 701–29.

McCallum, J. (1995). National borders matter: Canada–US regional trade patterns, *American Economic Review* 85: 615–23.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (ed.). *Frontiers in Econometrics*. New York: Academic Press, 105–42.

McLaren, J. (2000). 'Globalization' and vertical structure. *American Economic Review* 90: 1239–54.

McMillan, J., and Rothschild, M. (1994). Search. In: R.J. Aumann and S. Hart (eds.). *Handbook of Game Theory with Economic Applications, Volume 2*. Amsterdam: North-Holland, 905–27.

Meade, J.E. (1972). The theory of labour-managed firms and of profit-sharing. *Economic Journal* 62: 402–28.

Michel, Ph., Perrot, A., and Thisse, J.-F. (1996). Interregional equilibrium with heterogenous labor. *Journal of Population Economics* 9: 95–114.

Michel, Ph., Pestieau, P., and Thisse, J.-F. (1983). Regional allocation of investment with distributive objectives. *Journal of Regional Science* 23: 199–209.

Milgrom, P., and Roberts, J. (1992). *Economics, Organization and Management*. Englewood Cliffs, NJ: Prentice Hall.

Mills, E.S. (1967). An aggregative model of resource allocation in a metropolitan area. *American Economic Review* 57: 197–210.

Mills, E.S. (1970). The efficiency of spatial competition. *Papers and Proceedings of the Regional Science Association* 25: 71–82.

Mills, E.S. (1972a). *Studies in the Structure of the Urban Economy*. Baltimore: The Johns Hopkins Press.

Mills, E.S. (1972b). *Urban Economics*. Glenview, Ill: Scott, Foresman and Company.

Mirrlees, J. (1972). The optimum town. *Swedish Journal of Economics* 74: 114–35.

Mirrlees, J. (1995). Welfare economics and economies of scale. *Japanese Economic Review* 46: 38–62.

Miyagiwa, K. (1991). Scale economies in education and the brain drain problem. *International Economic Review* 32: 743–59.

Mori, T. (1997). A model of megapolis formation: the maturing of city systems. *Journal of Urban Economics* 42: 133–57.

Mougeot, M. (1978). The welfare foundations of regional planning: general equilibrium and Pareto optimality in a spatial economy. *Regional Science and Urban Economics* 8: 175–94.

Mossay, P. (2001). Market structure and migration in a spatial economy. Université catholique de Louvain, Department of Economics, Ph.D. thesis.

Mueller, D. (1989). *Public Choice II. A Revised Version of Public Choice.* Cambridge: Cambridge University Press.

Mulligan, G. (1984) Agglomeration and central place theory: a review of the literature. *International Regional Science Review* 9: 1–42.

Mussa, M. (1978). Dynamic adjustment in the Heckscher-Ohlin-Samuelson model. *Journal of Political Economy* 86: 775–91.

Muth, R.F. (1961). The spatial structure of the housing market. *Papers and Proceedings of the Regional Science Association* 7: 207–20.

Muth, R.F. (1969). *Cities and Housing*. Chicago: University of Chicago Press.

Myrdal, G. (1957). *Economic Theory and Underdeveloped Regions*. London: Duckworth.

Nagurney, A. (1993). *Network Economics: A Variational Inequality Approach*. Dordrecht: Kluwer Academic Publishers.

Neary, J.P. (2001). Of hype and hyperbolas: introducing the new economic geography. *Journal of Economic Literature* 49: 536–61.

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy* 78: 311–29.

Nerlove, M.L., and Sadka, E. (1991). Von Thünen's model of the dual economy. *Journal of Economics* 54: 97–123.

Netz, J.S., and Taylor, B.A. (2001). Maximum or minimum differentiation? An empirical investigation into the location of firms. *Review of Economics and Statistics*, forthcoming.

Norman, G., and Thisse, J.-F. (1996). Product variety and welfare under soft and tough pricing regimes. *Economic Journal* 106: 76–91.

Novshek, W., and Sonnenschein, H. (1987). General equilibrium with free entry. *Journal of Economic Literature* 25: 1281–306.

Ogawa, H., and Fujita, M. (1980). Equilibrium land use patterns in a non-monocentric city. *Journal of Regional Science* 20: 455–75.

Ogawa, H., and Fujita, M. (1985). Nonmonocentric urban configuration in a two-dimensional space. *Environment and Planning A* 21: 363–74.

O'Hara, D.J. (1977). Location of firms within a square central business district. *Journal of Political Economy* 85: 1189–207.

Ohlin, B. (1933). *Interregional and International Trade.* Cambridge, MA: Harvard University Press. Revised version published in 1968.

Ota, M., and Fujita, M. (1993). Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Regional Science and Urban Economics* 23: 695–729.

Ohmae, K. (1995). *The End of the Nation State*. New York: The Free Press.

Ottaviano, G.I.P. (1999). Integration, geography, and the burden of history. *Regional Science and Urban Economics* 29: 245–56.

Ottaviano, G.I.P (2001). Monopolistic competition, trade, and endogenous spatial fluctuations. *Regional Science and Urban Economics* 31: 51–77.

Ottaviano, G.I.P., and Puga, D. (1998). Agglomeration in the global economy: a survey of the 'new economic geography.' *World Economy* 21: 707–31.

Ottaviano G.I.P., Tabuchi, T., and Thisse, J.-F. (2002). Agglomeration and trade revisited. *International Economic Review*, forthcoming.

Ottaviano, G.I.P., and Thisse, J.-F. (1998). Agglomeration and trade revisited. CEPR Discussion Paper N°1903.

Ottaviano, G.I.P., and Thisse, J.-F. (2001). Integration, agglomeration and the political economics of factor mobility. *Journal of Public Economics*, forthcoming.

Papageorgiou, Y.Y. (1977). Fundamental problems of theoretical planning. *Environment and Planning A* 9: 1329–56.

Papageorgiou, Y.Y. (1983). Models of agglomeration. *Sistemi Urbani* 3: 391–410.

Papageorgiou, G.J., and Casetti, E. (1971). Spatial equilibrium residential land values in a muticentric city. *Journal of Regional Science* 11: 385–89.

Papageorgiou, Y.Y., and Pines, D. (1990). The logical foundations of urban economics are consistent. *Journal of Economic Theory* 50: 37–53.

Papageorgiou, Y.Y., and Pines, D. (1999). *An Essay in Urban Economic Theory*. Dordrecht: Kluwer Academic Publishers.

Papageorgiou, Y.Y., and Pines, D. (2000). Externalities, indivisibility, nonreplicability and agglomeration. *Journal of Urban Economics* 48: 509–35.

Papageorgiou, Y.Y., and Smith, T.R. (1983). Agglomeration as local instability of spatially uniform steady-states. *Econometrica* 51: 1109–19.

Papageorgiou, Y.Y., and Thisse, J.-F. (1985). Agglomeration as spatial interdependence between firms and households. *Journal of Economic Theory* 37: 19–31.

Peri, G. (2001). Human capital externalities in US cities. *Regional Science and Urban Economics*, forthcoming.

Perroux, F. (1955). Note sur la notion de pôle de croissance. *Economique appliquée* 7: 307–20.

Phlips, L. (1983). *The Economics of Price Discrimination*. Cambridge: Cambridge University Press.

Pirenne, H. (1925). *Medieval Cities*. Princeton: Princeton University Press.

Ponsard, C. (1983). *History of Spatial Economic Theory*. Heidelberg: Springer-Verlag.

Porter, M.E. (1998). *On Competition*. Cambridge, MA: A Harvard Business Review Book.

Pred, A. (1977). *City Systems in Advanced Economies*. New York: Halsted.

Prescott, E.C. (1998). Nedded: a theory of total factor productivity. *International Economic Review* 39: 525–51.

Puga, D. (1999). The rise and fall of regional inequalities. *European Economic Review* 43: 303–34.

Pyke, F., Becattini, G., and Sengenberger, W., eds. (1990). *Industrial Districts and Inter-firm Cooperation in Italy*. Geneva, International Institute for Labour Studies.

Quah, D.T. (1996). Empirics for economic growth and convergence. *European Economic Review* 40: 1353–75.

Quinzii, M., and Thisse, J.-F. (1990). On the optimality of central places. *Econometrica* 58: 1101–19.

Rauch, J. E. (1993). Productivity gains from geographical concentration of human capital: Evidence from the cities. *Journal of Urban Economics* 34: 380–400.

Razin A., and Sadka, E. (1997). International migration and international trade. In M.R. Rosenzweig and O. Stark (eds.). *Handbook of Population and Family Economics*. Amsterdam: North-Holland, 851–87.

Reilly, W.J. (1931). *The Law of Retail Gravitation*. New York: Pilsbury.

Richardson, H.W. (1987). The costs of urbanization: a four-country comparison. *Economic Development and Cultural Change* 35: 561–80.

Rivera-Batiz, F. (1988). Increasing returns, monopolistic competition, and agglomeration economies in consumption and production. *Regional Science and Urban Economics* 18: 125–53.

Roback, J. (1982). Wages, rents, and the quality of life. *Journal of Political Economy* 90: 1257–78.

Romer, P. (1986). Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–37.

Romer, P. (1990). Endogeneous technological change. *Journal of Political Economy* 98: S71–S102.

Romer, P. (1992). Increasing returns and new developments in the theory of growth. In: W.A. Barnett, B. Cornet, C. d'Aspremont, J.J. Gabszewicz, and A. Mas-Colell (eds.). *Equilibrium Theory with Applications*. Cambridge: Cambridge University Press, 83–110.

Rosen, S. (1979). Wage-based indexes of urban quality life. In: P. Mieszkowski and M. Straszheim (eds.). *Current Issues in Urban Economics*. Baltimore: John Hopkins University Press, 74–104.

Rosenstein-Rodan, P.N. (1943). Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53: 202–11.

Sachs J. (2000). A new map of the world, *The Economist* July 24th 2000, 99–101.

Sala-i-Martin, X. (1996). Regional cohesion: evidences and theories of regional growth and convergence. *European Economic Review* 40: 1325–52.

Salop, S.C. (1979). Monopolistic competition with an outside good. *Bell Journal of Economics* 10: 141–56.

Samuelson, P.A. (1939). The gains from international trade. *Canadian Journal of Economics* 5: 195–205.

Samuelson, P.A. (1952). Spatial price equilibrium and linear programming. *American Economic Review* 42: 283–303.

Samuelson, P.A. (1954a). The transfer problem and transport cost, II: analysis of effects of trade impediments. *Economic Journal* 64: 264–89.

Samuelson, P.A. (1954b). The pure theory of public expenditure. *Review of Economic Studies* 36: 387–9.

Samuelson, P.A. (1962). The gains form international trade once again. *Economic Journal* 72: 820–9.

Samuelson, P.A. (1983). Thünen at two hundred. *Journal of Economic Literature* 21: 1468–88.

Saxenian, A. (1994). *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Cambridge, MA: Harvard University Press.

Schulz, N., and Stahl, K. (1989). Good and bad competition in spatial markets for search goods: the case of linear utility function. *Annales d'Economie et de Statistique* 15/16: 113–36.

Schulz, N., and Stahl, K. (1996). Do consumers search for the highest price? equilibrium and monopolistic optimum in differentiated products markets. *Rand Journal of Economics* 27: 542–62.

Schweizer, U. (1978). A spatial version of the nonsubstitution theorem. *Journal of Economic Theory* 19: 307–20.

Schweizer, U. (1986). General equilibrium in space. In: J.J. Gabszewicz, J.-F. Thisse, M. Fujita, and U. Schweizer, *Location Theory*. Chur: Harwood Academic Publishers, 151–85.

Schweizer, U., and Varaiya, P.V. (1976). The spatial structure of production with a Leontief technology. *Regional Science and Urban Economics* 6: 231–51.

Schweizer, U., Varaiya, P.V., and Hartwick, J. (1976). General equilibrium and location theory. *Journal of Urban Economics* 3: 285–303.

Scitovsky, T. (1954). Two concepts of external economies. *Journal of Political Economy* 62: 143–51.

Scotchmer, S. (1986). Local public goods in an equilibrium: how pecuniary externalities matter, *Regional Science and Urban Economics* 16: 463–81.

Scotchmer, S. (1994). Public goods and the invisible hand. In: J. Quigley and E. Smolensky (eds.). *Modern Public Finance*. Cambridge, MA: Harvard University Press, 93–125.

Scotchmer, S. (2001). Local public goods and clubs. In: A. Auerbach and M. Feldstein (eds.). *Handbook of Public Economics, Volume 3*. Amsterdam: North-Holland, forthcoming.

Scotchmer, S., and Thisse, J.-F. (1999). Space and value: an outlook and new perspectives. In: L.-A. Gérard-Varet and A.P. Kirman (eds.). *Economics in the Next Millenium*. Oxford: Oxford University Press, 152–68.

Scotchmer, S., and Wooders, M. (1987). Competitive equilibrium and the core in club economies with anonymous crowding, *Journal of Public Economics* 34, 159–73.

Scott, A.J. (1976). Land use and commodity production. *Regional Science and Urban Economics* 6: 147–60.

Sen, A., and Smith, T.E. (1995). *Gravity Models of Spatial Interaction Behavior*. Heidelberg: Springer-Verlag.

Serk-Hanssen, J. (1969). The optimal number of factories in a spatial market. In: H.C. Bos (ed.). *Towards Balanced International Growth*. Amsterdam: North-Holland, 269–81.

Shields, G.M., and Shields, M.P. (1989). The emergence of migration theory and a suggested new direction. *Journal of Economic Surveys* 3: 277–304.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*, Straham and Cadell, London. Fifth edition reprinted in 1965. New York. The Modern Library.

Smith, W.D. (1984). The function of commercial centers in the modernization of European capitalism: Amsterdam as an information exchange in the seventeenth century. *Journal of Economic History* 44: 985–1005.

Smith, P.J. (1999). Do knowledge spillovers contribute to U.S. state output and growth? *Journal of Urban Economics* 45: 331–53.

Solow, R.M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.

Solow, R.M. (1973). On equilibrium models of urban locations. In: J.M. Parkin (ed.). *Essays in Modern Economics*. London: Longman, 2–16.

Sonnenschein, H. (1982). Price dynamics based on the adjustment of firms. *American Economic Review* 72: 1088–96.

SOPEMI (1998). *Trends in International Migration*. Paris: OECD.

Soubeyran, A., and Weber, S. (2002). District formation: a co-opetition approach, *Journal of Urban Economics*, forthcoming.

Spence, M. (1976). Product selection, fixed costs, and monopolistic competition. *Review of Economic Studies* 43: 217–35.

Sraffa, P. (1926). The laws of return under competitive conditions. *Economic Journal* 36: 535–50.

Stahl, K. (1982). Differentiated products, consumer search, and locational oligopoly. *Journal of Industrial Economics* 31: 97–114.

Stahl, K. (1983). A note on the microeconomics of migration. *Journal of Urban Economics* 14: 318–26.

Stahl, K. (1987). Theories of urban business location. In: E.S. Mills (ed.). *Handbook of Regional and Urban Economics, Volume 2*. Amsterdam: North-Holland, 759–820.

Starrett, D. (1974). Principles of optimal location in a large homogeneous area. *Journal of Economic Theory* 9: 418–48.
Starrett, D. (1978). Market allocations of location choice in a model with free mobility. *Journal of Economic Theory* 17: 21–37.
Starrett, D. (1988). *Foundations of Public Economics*. Cambridge: Cambridge University Press.
Stern, N. (1972). The optimal size of market areas. *Journal of Economic Theory* 4: 154–73.
Stigler, G.J. (1961). The economics of information. *Journal of Political Economy* 69: 213–25.
Stiglitz, J. (1977). The theory of local public goods. In: M.S. Feldstein and R.P. Inman (eds.). *The Economics of Public Services*. London: Macmillan, 273–334.
Stollsteimer, J.F. (1963). A working model for plant numbers and locations. *Journal of Farm Economics* 45: 631–45.
Stuart, C. (1979). Search and the spatial organization of trading. In: S. Lipman and J.J. McCall (eds.). *Studies in the Economics of Search*. Amsterdam: North-Holland, 17–33.
Tabuchi, T. (1986). Urban agglomeration economies in a linear city. *Regional Science and Urban Economics* 16: 421–36.
Tabuchi, T. (1998). Agglomeration and dispersion: a synthesis of Alonso and Krugman. *Journal of Urban Economics* 44: 333–51.
Takayama, T., and Judge, G.G. (1971). *Spatial and Temporal Price and Allocation Models*. Amsterdam: North-Holland.
Tauchen, H., and Witte, A.D. (1984). Social optimal and equilibrium distribution of office activity: models with exogenous and endogenous contacts. *Journal of Urban Economics* 15: 242–61.
Teece, D.J. (1977). Technology transfer by multinational firms: the resource cost of transferring technological know-how. *Economic Journal* 87: 242–61.
Teitz, M.B. (1968). Towards a theory of urban facility location. *Papers and Proceedings of the Regional Science Association* 21: 35–52.
Thill, J.-C., and Thomas, I. (1987). Toward conceptualizing trip chaining: a review. *Geographical Analysis* 19: 1–17.
Thisse, J.-F., and van Ypersele, T. (1999). The challenge raised by metropolisation and fiscal competition in economic development. *The World Economy* 22: 1201–20.
Thisse, J.-F., and Vives, X. (1988). On the strategic choice of a spatial price policy. *American Economic Review* 78: 122–37.
Thisse, J.-F., and Zoller, H.G. (1983). *Locational Analysis of Public Facilities*. Amsterdam: North-Holland.
Thünen, J.H. von (1826). *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Hamburg: Perthes. English translation: *The Isolated State*. Oxford: Pergammon Press (1966).
Tiebout, C.M. (1956). A pure theory of local public expenditures. *Journal of Political Economy* 64: 416–24.
Tiebout, C.M. (1961). An economic theory of fiscal decentralization. In: NBER, *Public Finances: Needs, Sources and Utilization*. Princeton, NJ: Princeton University Press.
Tinbergen, J. (1962). *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: Twentieth Century Fund.

Toffler, A. (1980). *The Third Wave*. New York: Bantam Books.
United Nations (1994). *World Urbanization Prospects: The 1992 Revision*. New York: United Nations, Population Division.
Varga, A. (2000). Local academic knowledge transfers and the concentration of economic activity. *Journal of Regional Science* 40: 298–309.
Venables, A.J. (1996). Equilibrium locations of vertically linked industries. *International Economic Review* 37: 341–59.
Vickrey, W. (1964). *Microstatics*. New York: Harcourt, Brace and World.
Vickrey, W. (1977). The city as a firm. In: M.S. Feldstein and R.P. Inman (eds.). *The Economics of Public Services*. London: Macmillan, 334–43.
Vives, X. (1990). Trade association disclosure rules, incentives to share information, and welfare, *Rand Journal of Economics* 21: 409–30.
Waltz, U. (1996). Transport costs, intermediate goods, and localized growth. *Regional Science and Urban Economics* 26: 671–95.
Webber, M.J. (1979). *Information Theory and Urban Spatial Structure*. London: Croom Helm.
Weber, A. (1909). *Über den Standort der Industrien.* Tübingen: J.C.B. Mohr. English translation: *The Theory of the Location of Industries.* Chicago: Chicago University Press, 1929.
Wesolowosky, G.O. (1993). The Weber problem: history and perspectives. *Location Science* 1: 5–23.
Wheaton, W.C. (1974). A comparative static analysis of urban spatial structure. *Journal of Economic Theory* 9: 223–37.
Wheaton, W.C. (1977). Income and urban residence: an analysis of consumer demand for location. *American Economic Review* 67: 620–31.
Whitaker, J.K. (1998). Henry George on the location of economic activity. In: M. Bellet and C. L'Harmet (eds.). *Industry, Space and Competition*. Cheltenham: Edward Elgar. 174–84.
Wildasin, D. (1986a). Spatial variation of marginal utility of income and unequal treatment of equals. *Journal of Urban Economics* 19: 125–29.
Wildasin, D.E. (1986b). *Urban Public Finance*. Chur: Harwood Academic Publishers.
Wildasin, D. (1987). Theoretical analysis of local public economics. In: E.S. Mills (ed.). *Handbook of Regional and Urban Economics, Volume 2*. Amsterdam: North-Holland, 1131–78.
Wolf, H.C. (2000). Intranational home bias in trade. *Review of Economics and Statistics* 82: 555–63.
Wolinsky, A. (1983). Retail trade concentration due to consumers' imperfect information. *Bell Journal of Economics* 14: 275–82.
World Bank (2000). *Entering the 21st Century. World Development Report 1999/2000*. Oxford: Oxford University Press.
Young, A. (1928). Increasing returns and economic progress. *Economic Journal* 38: 527–42.

Subject Index