Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

Fabio Crescenzi
Stefania Mignani *Editors*

Statistical Methods and Applications from a Historical Perspective

Selected Issues

Springer

Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

For further volumes:
http://www.springer.com/series/10104

Series Editors

Spanish Society of Statistics and Operations Research (SEIO)

Société Française de Statistique (SFdS)

Società Italiana di Statistica (SIS)

Sociedade Portuguesa de Estatística (SPE)

Fabio Crescenzi • Stefania Mignani
Editors

Statistical Methods and Applications from a Historical Perspective

Selected Issues

Springer

Editors
Fabio Crescenzi
Istat Head of Office for Census Methods
National Institute of Statistics
Rome
Italy

Stefania Mignani
Department of Statistical Sciences
University of Bologna
Bologna
Italy

ISSN 2194-7767 ISSN 2194-7775 (electronic)
ISBN 978-3-319-05551-0 ISBN 978-3-319-05552-7 (eBook)
DOI 10.1007/978-3-319-05552-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014941586

Mathematics Subject Classification (2010): 62A, 62F, 62H, 91B, 91C, 91D

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Dear reader, On behalf of the four Scientific Statistical Societies – the SEIO, Sociedad de Estadística e Investigación Operativa (Spanish Statistical Society and Operation Research); SFdS, Société Française de Statistique (French Statistical Society); SIS, Societá Italiana di Statistica (Italian Statistical Society); and the SPE, Sociedade Portuguesa de Estatística (Portuguese Statistical Society) – we would like to inform you that this is a new book series of Springer entitled "Studies in Theoretical and Applied Statistics," with two lines of books published in the series: "Advanced Studies" and "Selected Papers of the Statistical Societies."

The first line of books offers constant up-to-date information on the most recent developments and methods in the fields of theoretical statistics, applied statistics, and demography. Books in this series are solicited in constant cooperation between the statistical societies and need to show a high-level authorship formed by a team preferably from different groups so as to integrate different research perspectives.

The second line of books presents a fully peer-reviewed selection of papers on specific relevant topics organized by the editors, also on the occasion of conferences, to show their research directions and developments in important topics, quickly and informally, but with a high level of quality. The explicit aim is to summarize and communicate current knowledge in an accessible way. This line of books will not include conference proceedings and will strive to become a premier communication medium in the scientific statistical community by receiving an Impact Factor, as have other book series such as "Lecture Notes in Mathematics."

The volumes of selected papers from the statistical societies will cover a broad range of theoretical, methodological as well as application-oriented articles, surveys and discussions. A major goal is to show the intensive interplay between various, seemingly unrelated domains and to foster the cooperation between scientists in different fields by offering well-founded and innovative solutions to urgent practice-related problems.

On behalf of the founding statistical societies I wish to thank Springer, Heidelberg and in particular Dr. Martina Bihn for the help and constant cooperation in the organization of this new and innovative book series.

Rome, Italy — Maurizio Vichi

Preface

The celebration of the 150th anniversary of the unification of Italy was accompanied by several initiatives of great interest for the scientific community.

The socio-economic changes which have taken place since 1861 have provided an opportunity to reflect in a qualified and very broad cultural context on how Italy has gone through these years of great changes, even considering the most important developments in other European countries.

The information collected by public institutions such as Istat and Bank of Italy has increased over time in terms of quantity, availability and, above all, reliability. The analysis of these data allows us to understand the dynamics of the real phenomena and consequently gives us tools to better evaluate the right decision-making strategies for the development of the country.

The study of economic and social phenomena evolution must be put in connection to the evolution of the methods and tools used in each historical phase trying to determine the suitability of statistical methods and tools employed for data analysis. The study of the past to understand the present and plan the future is a challenge for the research and necessarily involves also a critical analysis of the interaction among phenomena, methods and tools employed to investigate on them using an appropriate statistical approach.

Considering this background, the Italian Statistical Society in collaboration with Istat and Bank of Italy organized the Biennial Conference of Statistics dedicated to the analysis of economic phenomena, demographic and social changes that have affected our country since its Unification. Past and current developments of methodological issues and their applications have been dealt in an historical comparative analysis to investigate on future prospects.

This volume contains a selection of peer-reviewed papers, whose preliminary version was presented at this meeting.

The meeting was an opportunity that brought together experts, from Italy and other European countries, with more than 250 participants. It was an important opportunity to discuss crucial methodological issues and its critical applications, not only from an historical perspective but also considering future developments.

This volume is divided into the following six parts:

- Methodological issues in an historical perspective
- Historical issues in economic and socio-demographic studies
- New developments in survey methodology for official statistics
- New methodological developments in economic studies

- New methodological developments in educational studies
- New methodological developments in social and demographic studies

Each part includes a number of papers ranging from 5 to 7 selected after a meticulous process of review which involved a larger number of papers. We wish to thank especially all the authors and the reviewers implicated in this demanding task. Besides we would like to thank Dr. A. Blank from Springer-Verlag for the excellent cooperation provided in the whole process of publication of this volume.

In conclusion, we would like to emphasize once again the relevance of the debate that we had on the wide range of different issues covered in this volume, each of which will be crucial for improving the methods to analyze changes and to support the development of the country in the coming years.

Rome, Italy — Fabio Crescenzi
Bologna, Italy — Stefania Mignani

Contents

Part I

Methodological Issues in an Historical Perspective

Bayesian and Non-Bayesian Approaches to Statistical Inference: A Personal View

Bruno Chiandotto

Abstract

Bayesian and non-bayesian approaches to statistical inference are compared giving particular attention to the emerging field of causal statistical inference and causal statistical decision theory. After a brief review of the evolution of statistical inference, as extraction of information and identification of models from data, the problematic issues of causal inference and causal decision theory will be reviewed. The aim is to provide some basic ideas for unifying the different approaches and for strengthening the future of statistics as a discipline.

Prologue

Hume (2003) argued that induction is irrational. This view, often called Humean irrationalism, conflicts with the empiricist view that affirms that science proceeds in a rational and inductive way. Many attempts have been made to refute Hume. One of the earliest is due to Bayes (1763) and Laplace (1812). According to Bayes, rational learning proceeds by assigning probabilities Keynes (1921), usually called prior probabilities, to hypotheses. Using Bayes's theorem, these prior probabilities are then updated in the light of experience. To determine these probabilities, Laplace used what is often called the *principle of insufficient reason*.

Subsequently, the Laplacian account of rational learning was criticized as applying the same intuition to a different representation of the problem often yields different probabilities. Keynes (1921) and Carnap (1950) tried to improve Laplace's approach by interpreting the prior probabilities as a measure of quantifying logical relations between statements. Fisher (1930, 1935, 1956) and Popper (1959) sharply

B. Chiandotto (✉)
Department of Statistics, University of Florence, Florence, Italy
e-mail: chiandot@ds.unifi.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7__1,

rejected the Bayes–Laplace tradition and proposed other solutions to the problem of rational learning. With his theory of significance testing, Fisher revolutionized statistical theory and practice. Meanwhile, Popper developed the falsificationist methodology and had a similar influence on the philosophy of science. Both solutions to the problem of rational learning are based on the same principle, namely, that it is rational to accept hypotheses if they have survived rigorous testing. In Popper's terminology, such hypotheses are called corroborated. A similar approach is due to Gini (1943)[1] and Pompilj (1952, 1961).

1 Introduction

The history of statistical inference is marked by controversies about its fundamental principles. Historically, one can consider roughly four principal approaches to statistical inference.

The first approach is called Fisherian. Fisher has emphasized the need for a variety of approaches for different problems; he was dismissive of axiomatic arguments. A second approach due to Neyman and Pearson (1928), initially developed to explain Fisher's ideas more concretely, is strongly based on the frequency theory of probability and emphasize operational concepts. A third approach, where probability represents a rational degree of belief, in which different people faced by the same evidence share the same probability, goes back to Laplace and his predecessors and in its modern form it is associated with Carnap and, especially, with Jeffreys (1931). This (*objective*) approach has been extended by specific characterization of probability in which the degree of belief is constrained only by the requirement of self-consistency. In this fourth approach, (*personalistic or subjective*), associated with Ramsey (1931), Good (1960), De Finetti (1937) and Savage (1951, 1954), there is no assumption that different people with the same knowledge express the same probability on a specified event.

In the first two approaches, usually referred to as *classical theory of statistical inference*, the procedures are justified by their performance under hypothetical repetitions of the experiment, i.e. frequency properties. The differences between the two are minor and are essentially the following: (a) in the Fisherian approach, emphasis is placed on the simple test of significance, on the likelihood function and principles as sufficiency; (b) in the Neyman–Pearson approach, operational requirements, such as power and other explicit indicators of sensitivity, are emphasized and confidence interval and acceptance and rejection of hypotheses terminology are introduced.

Jeffreys's approach to inference has the same target as Fisher's: what can be reasonably learned about a parameter of the hypothesized model from the data? But, in contrast to Fisher, Jeffreys argues that a different notion of probability is needed to achieve this, specifically, a reasonable degree of belief computed by means of

[1] On the contributions of Gini to the foundations of probability and statistical inference I strongly recommend a forthcoming paper by Piccinato (2011).

Bayes's rule; the a priori distribution is taken, in accordance with Laplace, to be dispersed, representing lack of knowledge.

Jeffreys's and the personalistic approach are often referred to as *Bayesian* (or *neo-bayesian*) *approaches to statistical inference*. Although they are formally the same, there are some fundamental and philosophical differences: the personalistic degree of belief, in contrast to the reasonable degree of belief, measures how strongly you believe in something in the light of the model for the data; the direct consequence is that the choice of the prior is substantially different.

There are other approaches to statistical inference. The most relevant are: fiducial inference, likelihood inference, plausibility inference, structural inference, pivotal inference, prequential inference, and predictive inference.

All the approaches to statistical inference utilize some kind of information to obtain a description (through a statistical model) of the phenomenon under study. In my view, every approach based on mathematical models should accommodate all the different approaches and provide tools for making comparative analyses. Such an approach is the *decision approach* substantially already present in the Fisher and Neyman–Pearson theories. Moreover, the decision approach gives a satisfactory solution so far, at the so-called *pragmatic problem of induction*.

Many authors (Cox 1958; Smith 1961) affirm that a distinction must be made between statistical inference and statistical decision theory. But other authors such as Lindley (2006), and this is also my opinion, consider statistical decision theory as one of the possible extensions of statistical inference. Moreover, the decision approach, combining various theories of statistical inference, avoids dogmatisms that can lead to paradoxical situations. It is free from logical error, is more effective in applications, and treats successfully a broader range of problems than competing approaches.

2 Bayesianism

Even if the most influential version of neo-bayesianism has been proposed by Savage, the term Bayesianism is used in a wider sense than Savage's approach. It includes the logical probability, frequentist probability, and some other attempts to objectify prior probabilities. Savage showed that a reasonable preference order over the set of all conceivable strategies can be represented by expected utilities of strategies, where now not only the utilities but also the probabilities for computing the expectations can be derived from the preference order. Substantially, Savage provided a general theory of rational learning and decision making. The relevance of neo-bayesianism, where all probabilities are the subjective degrees of belief, lies in the fact that it is a very general philosophy that seamlessly covers science and decision making starting from the problem of induction. Bayesian rationality constitutes progress beyond Humean irrationalism. Even if Bayesianism is not helpful when nothing is known, it might be helpful in the case of partial rather than complete prior information (Joyce 2010).

Real-world decision problems often have to be simplified to become tractable. According to contemporary model-building wisdom, *finding the right simplifications is an art, not a science*; it involves knowledge and requires experts in the field, this conviction is widely shared by experts. Bayesianism, it seems, gives the experts a possibility to bring their experience to bear on the problem. They can choose a prior probability measure in the light of their experience. Given this choice, which can be communicated to others, decision making can proceed, if the computations are feasible; if not, one can try to find an approximation. Indeed, model building is itself a matter of approximation; Bayesian experts might construct simplified models by excluding possibilities that they assign, in the light of their experience, a low prior probability. Thus, it could be argued that Bayesianism describes a rational way of expressing partial expert knowledge that cannot easily be expressed in another way. However, Bayesianism leaves in the dark how experts proceed when trying to transform experience into a prior. On the other hand, experts might learn from experience in a rational fashion. In this case, we already know how ideal Bayesian experts proceed. They start with a prior probability before making experiences, updating their prior, and when after some time they are viewed as experts, the prior they bring to a new problem is actually a posterior probability measure embodying their experience. The problem with this analysis is, however, that the everything-goes theorem implies that the expert's posterior is arbitrary. According to Bayesianism, all conclusions drawn from experience are equally reasonable or unreasonable.

3 Decision Theory and Utility

The foundations of the (*normative*) modern *statistical decision theory* is due to Von Neumann and Morgenstern (1947), for the so-called Expected Utility (EU) and Savage for the Subjective Expected Utility (SEU). These authors, on the basis of a series of postulates, or rational axioms of behavior of the decision maker, prove the existence of a real-valued utility function that can be derived from the betting rule.

Decision theory recommends an act that maximizes utility, that is, an act whose utility equals or exceeds the utility of every other act. It evaluates an act utility by calculating the act expected utility. It uses probabilities and utilities of an act possible outcomes to define an act expected utility.

Since people usually do not behave in ways consistent with the *axiomatic* rules and hence lead to violations of optimality, there is a related area of study, called a *descriptive decision theory*, attempting to describe what people actually do.

A series of criticisms (particularly Allais 1953 and, for an up-to-date and reasonably extended review, Chiandotto and Bacci 2004) have been made against EU and SEU. The criticisms regards, mostly, the empirical relevance of the rational axioms of behavior.

Even if the problem of the importance of the axioms on the behavior of the decision maker has to be viewed not in the sense of a good description, but in that

of a good rule (i.e., it concerns identifying the best decision to take, assuming an ideal decision maker who is fully informed, able to compute with perfect accuracy and full rationality) different authors have proposed alternative systems of axioms less restrictive and more compatible with the actual behavior of decision makers.

To generalize the normative decision theory, some authors adopted different terminology like *prescriptive decision theory* (Bell et al. 1988), *constructive decision theory* (Roy 1993; Tsoukiàs 2007). These approaches hypothesize weaker axioms than the classic ones; in particular, since the more frequently violated axiom is independence, the new theories release the property of linearity in the probability. Machina (1982) develops a utility theory without the presence of the independence axiom. Other theories, instead, do not include the axiom of transitivity (Fishburn 1973). Among the more interesting theoretical proposals (generalization of utility theories) we should include the rank-dependent utility (Quiggin 1993), the prospect theory (Kahneman and Tversky 1979), and cumulative prospect theory (Tversky and Kahneman 1992). Aiming at giving to decision theory useful operating tools, it must be considered the so-called causal approach to the theory of the decisions. This approach, although mainly developed in the context of the philosophical reflection, results of large interest for his statistical implications.

4 Causality

In spite of the innumerable developments, generalized utility theories are still not able to solve in a satisfactory way operative decision-making problems. In fact such theories discuss situations in which the consequences of acts are dependent on the *state of the world* whenever the action chosen has no effect on such state. This hypothesis in many contexts is not satisfied. In fact, in many situations the choice made by the decision maker has a, sometime, relevant effect on the state of nature (*the act causes the state*). Therefore, to solve decision problems, the analysis of causality becomes relevant in its theoretical aspects and in its operative implications.

Regarding causality, the paper of Freedman (1999) and three contributions of Mealli et al. (2011), Cox and Wermuth (2004), and of Frosini (2006) are especially useful. This latter author presents a synthetic but exhaustive panorama of the developments of the concept of causality: starting from the Aristotelian doctrine of causation he arrives to the more recent developments on relevant aspects to statistical modeling and, particularly, on acyclic graphical models (*Directed Acyclic Graphs—DAGs*). Also Cox and Wermuth, after an interesting close examination of three different definitions of causality, analyze graphical models focusing on the concepts of statistical independence and particularly on the difference between conditioning and intervention.

The paper of Mealli, Pacini, and Rubin gives a complete and up-to-date account of the so-called Neyman–Rubin–Holland model of causality. The framework proposed especially by (Rubin 1974, 2004; Holland and Rubin 1988) is very powerful and general, it provides a definition of causal effects in terms of potential outcomes, as well as a general statement of the assumptions, sufficient to make

causal inferences possible, even with observational data. Unfortunately, because of its generality the standard Neyman–Rubin–Holland model operates at a level of abstraction that is far away from the underlying mechanisms and processes that account for how observational data are generated. While such generality makes the model very powerful, its agnosticism about the underlying causal mechanisms can make it difficult to be applied in settings that are not close to a well-designed experiment.

Graphical models (Lauritzen and Richardson 2002) represent a generalization of the graphs of influence (Howard and Matheson 1984; Dawid 2002) that represent an extension of the path diagrams proposed by Wright (1921). In path analysis, the connections among the variables of interest are expressed in a graphical form, allowing to distinguish spurious from causal, direct and indirect effects, of variables. Other very interesting contributions to the statistical analysis of causality are Dawid (2000); Holland (1986); Pearl (1995, 2009); Spirtes et al. (2000) and, above all, Woodword (2003).

Woodword collects in his volume a 30-year of research activity presenting a new theory of causality that he considers superior to the counterfactual theory of causality developed by Lewis. The contribution of Woodword is placed in line with the studies of Spirtes, Glymour, and Scheines and of Pearl. While these latter authors concentrate their attention on the theoretical–methodological aspects, Woodword deals particularly with the philosophical foundations of the reasoning introducing a simple, but clear, definition of causality: C causes A if and only if the value of A is modified by an intervention on C. Woodword presents the tools for the analysis, graphics, and equations, for proceeding to the development of its *theory of manipulation*.

The different approaches to causality outlined above are characterized by specificities that are considered by the authors themselves not compatible: each author considers his own approach to be superior to the others. In my opinion, this position does not appear acceptable, as many of them are compatible at least in some fundamental aspects. Regarding superiority, there does not exist a statistical tool of universal validity able to give a satisfactory solutions in all research frameworks. The combined use of different approaches (Lauritzen 2004; White and Chalak 2006) seems the correct route to pursue for achieving the more interesting and significant results.

5 Causal Decision Theory

How much what we have said about causality can be relevant in the decision-making context? Causal decision theory adopts principles of rational choice that attend to an act consequences. It maintains that an account of rational choice must use causality to identify the considerations that make a choice rational. An act expected utility is a probability-weighted average of its possible outcome utilities. Possible states of the world that are mutually exclusive and jointly exhaustive, and so form a partition, generate an act possible outcomes. An act-state pair specifies an

outcome. Each product specifies the probability and utility of a possible outcome. The sum is a probability-weighted average of the possible outcomes utilities, where the probabilities depend casually on the act, probability are causal rather than merely evidential.

Joyce (1999) gives an account of rational decision making and probabilistic theories of evidence and confirmation. This author begins with an historical introduction to the topic of decision theory, including a critical discussion of Savage's theory, followed by a treatment of the modern *evidential theory* of decision making. Two chapters are deal with causal decision theory. The final chapter reports a unified representation theorem that simultaneously provides a firm foundation for both evidential and causal decision theory.

The accounts of rational decision discussed by Joyce presuppose that a rational agent should act so as to maximize some sort of "expected utility," which is a sort of weighted average of the utilities of the outcomes of a decision. What's at issue in the foundational disputes is which kind of expected utility should be maximized, and, consequently, which weights should be used in the weighted average of the values of the outcomes. All parties seem to agree that the weights should be set according to the probabilities of the outcomes given that the act is performed. The disagreement concerns how to unpack this subtle conditional-like expression for the purpose at hand. Evidential decision theory recommends performing that act which provides the *best evidence* for the good outcomes (on average). On the other hand, causal decision theorists propose a different way of unpacking. They suggest that we unpack this as the degree to which the act causally promotes the state. Several interpretations of causal probability have been proposed in the literature, and the connections between the various kinds of conditionals have been studied extensively in recent decades.

Armendt (1986), in a paper on the foundations of causal decision theory, distinguishes three different approaches to causal decision theory, similar in the contents but philosophically different, that go back, respectively, to Gibbard and Harper (1976), Skyrms (1979) and Lewis (1981). Gibbard and Harper distinguished causal decision theory, which uses probabilities of subjunctive conditionals, from evidential decision theory, which uses conditional probabilities. As in decision problems probabilities of subjunctive conditionals track causal relations, using them to calculate an option expected utility makes decision theory causal. They argued that expected utility, calculated with probabilities of conditionals, yields genuine expected utility. Skyrms presented a version of causal decision theory that dispenses with probabilities of subjunctive conditionals. His theory separates factors that the agent's act may influence from factors that the agent's act may not influence. Lewis defines the expected utility of an option and his formula for an option expected utility that is the same as Skyrms. The handy interpretation of the probability of a state if one performs an act, however, is not completely satisfactory. A good decision aims to produce a good outcome rather than evidence of a good outcome. Causal decision theory interprets the probability of a state, if one performs an act, as a certain type of causal probability rather than as a standard conditional probability. This aspect makes expected utility track efficacy, rather than auspiciousness.

As already outlined, Pearl, Spirtes, Glymour, and Scheines and Woodward present methods of inferring causal relations from statistical data. They use DAGs and associated probability distributions to construct causal models. In a decision problem, a causal model yields a way of calculating an act effect. A causal graph and its probability distribution express a dependency hypothesis and yield each act causal influence given that hypothesis. They specify the causal probability of a state under supposition of an act. An act expected utility is a probability-weighted average of its utilities according to the dependency hypotheses that candidate causal models represent.

Heckerman and Shachter (1995) proposed a version of Pearl's causality definition in the decision-making framework. This formulation has been rejected by Pearl himself. Heckerman and Shachter (2003) some years later, discussing the work "Statistics and Causal Inference" of Pearl, say: "... Unfortunately, Pearl has downplayed the strong connections between his work and decision theory as well as the suitability of the influence diagram as a representation of causal interactions. On the contrary, we believe that people who are familiar with decision theory will find comfort, as we have, in these connections ...."

6 Conclusions

The importance of Bayesianism, in which all probabilities are subjective degrees of belief, lies in the fact that it is a very general philosophy that seamlessly covers science and decision making from the problem of induction, which provides the context where it originated, to the theoretical and practical problems of statistical inference. Bayesian rationality constitutes a progress beyond Humean irrationalism. Even if Bayesianism is not helpful when nothing is known, it might be helpful in the case of partial rather than complete knowledge (Joyce 2010).

Causal knowledge plays an important role in everyday reasoning, it enables to predict future outcomes, explain past events, control the environment. Correlations among events can often be good indicators of the presence of some causal relation, but it is well known that observed associations are insufficient to disambiguate causal structure. For this reason much of causal learning takes place in the context of intervention that, in the real world often involves learning a complex network of relations among many events (Pearl 2011). To learn from interventions one must first decide which intervention to make.

Intervention is the central subject of the contributions of Pearl on causality. This author, in my opinion, has given the more interesting and innovative contributions to the analysis of causality, but his contributions, to become really useful from an empirical point of view, must be reinterpreted, as suggested by Heckerman and Schachter, in a decision theoretic framework. The decision-making process allows learners to use interventions to disambiguate particular causal structures, namely, those that they have in mind as potential models of the causal system.

References

Allais, M.: Le comportament de l'homme rationel devant le risque: critique des axiom et postulates de l'ecole americane. Econometrica **21**, 503–546 (1953)

Armendt, B.: A foundations of causal decision theory. Topoi **5**, 3–19 (1986)

Bayes, T.: Essay towards solving a problem in the doctrine of chances. Philos. Trans. R. Soc. Lond. **53**, 370–418 (1763)

Bell, D.E., Raiffa, H., Tversky, A.: Decision Making. Cambridge University Press, Cambridge (1988)

Carnap, R.: Logical Foundations of Probability. Chicago University Press, Chicago (1950)

Chiandotto, B., Bacci, S.: Decisioni razionali per il governo dell'università, un prerequisito essenziale: la teoria dell'utilità. Università degli Studi di Firenze (2004)

Cox, D.R.: Some problems connected with statistical inference. Ann. Math. Stat. **29**, 357–372 (1958)

Cox, D.R., Wermuth, N.: Causality: a statistical view. Int. Stat. Rev. **72**, 285–305 (2004)

Dawid, A.P.: Causal inference without counterfactuals. J. Am. Stat. Assoc. **95**, 407–448 (2000)

Dawid, A.P.: Influence diagrams for causal modelling and inference. Int. Stat. Rev. **70**, 161–189 (2002)

De Finetti, B.: La prévision: ses lois logiques, ses source subjectives. Ann. del'Institut Henri Poincarré **24**, 17–24 (1937)

Fishburn, P.: A mixture-set axiomatization of conditional subjective expected utility. Econometrica **41**, 1–25 (1973)

Fisher, R.A.: Inverse probability. Math Proc. Camb. Philos. Soc. **26**, 528–535 (1930)

Fisher, R.A.: The logic of Inductive inference (with discussion). J. R. Stat. Soc. **98**, 39–82 (1935)

Fisher, R.A.: Statistical Method and Scientific Inference. Oliver and Boyd, Edinburgh (1956)

Freedman, D.: From association to causation: some remarks on the history of statistics. Stat. Sci. **14**, 243–258 (1999)

Frosini, B.V.: Causality and causal models: a conceptual perspective. Int. Stat. Rev. **74**, 305–334 (2006)

Gibbard, A., Harper, W.: Counterfactuals and two kinds of expected utility. In: Harper, W., Stalnaker, R., Pearce, G. (eds.) Conditionals, Belief, Decision, Chance, and Time, pp. 153–190. Dordrecht-Reidel, Dordrecht (1976)

Gini, C.: I test di significatività, Atti della VII Riunione della Società Italiana di Statistica, Roma (1943)

Good, I.J.: Weights of evidence, corroboration, explanatory power, information and utility of experiment. J. R. Stat. Soc. Ser. B Stat. Methodol. **22**, 319–331 (1960)

Heckerman, D., Shachter, R.: Decision-theoretic foundations for causal reasoning. J. Artif. Intell. Res. **3**, 405–430 (1995)

Heckerman, D., Shachter, R.: Discussion in Pearl, J.: Statistics and causal inference: a review. Test **12**, 101–165 (2003)

Holland, P.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–960 (1986)

Holland, P., Rubin, D.: Causal inference in retrospective studies. Eval. Rev. **13**, 203–231 (1988)

Howard, R.A., Matheson, J.E.: Influence diagrams. In: Howard, R.A., Matheson, J.E. (eds.) Readings in the Principles and Applications of Decision Analysis. Strategic Decision Group, Menlo Park (1984)

Hume, D.: A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects. Book 1: Of the Understanding (2003). The Project Gutenberg EBook. Release Date: December, 2003

Jeffreys, H.: Scientific Inference. Cambridge University Press, Cambridge (1931)

Joyce, J.: The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge (1999)

Joyce, J.: A defense of imprecise credence and decision making. Philos. Perspect. **24**, 281–323 (2010)

Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**, 263–293 (1979)
Keynes, J.M.: A Treatise on Robability. MacMillan & Co. London (1921)
Laplace, P.S.: Théorie analitique the probabilities, 3rd edn. (1820). Courcier, Paris (1812)
Lauritzen, S.L.: Discussion on causality in Rubin, D.B. (2004): Direct and indirect causal effects via potential outcomes. Scand. J. Stat. **31**, 161–170 (2004)
Lauritzen, S., Richardson, T.: Chain graph models and their causal interpretations (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. **64**, 321–361 (2002)
Lewis, D.: Causal decision theory. Australas. J. Philos. **59**, 5–30 (1981)
Lindley, D.V.: Understanding Uncertainty. Wiley, Hoboken (2006)
Machina, M.J.: Expected utility analysis without the independence axiom. Econometrica **50**, 227–323 (1982)
Mealli, F., Pacini, B., Rubin, D.B.: Statistical inference for causal effects. In: Kenett, R., Salini, S. (eds.) Modern Analysis of Customer Satisfaction Surveys. Wiley, Chichester (2011)
Neyman, J., Pearson, E.S.: On the use and interpretation of certain test criteria for purposes of statistical inference. Part. I and II. Biometrika **20A**, 175–240 and 263–294 (1928)
Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–710 (1995)
Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, Cambridge (2009)
Pearl, J.: The causal foundations of structural equation modeling. In: Hoyle, R.H. (ed) Handbook of Structural Equation Modeling. Guilford Press, New York (2011)
Piccinato, L.: Gini's criticism to the theory of inference: a missed opportunity. Metron **69**, 101–117 (2011)
Pompilj, G.: Logica della conformità. Archimede **4**, 22–28 (1952)
Pompilj, G.: Teoria dei campioni, Applicazioni alla sperimentazione, alla produttività e alle rilevazioni campionarie. Partial reprint (1956), Veschi, Roma (1961)
Popper, K.R.: The Logic of Scientific Discovery. First version of this book appeared as Logic der Forschung, 1934. Hutchison Education, London (1959)
Quiggin, J.: Generalized Expected Utility the Rank-Dependent Model. Kluwer, Boston (1993)
Ramsey, F.P.: Truth and probability. In: Braithwaite, R.B. (ed.) The Foundations of Mathematics and Other Logical Essays, pp. 56–198. Routledge & Kegan Paul, London (1931)
Roy, B.: Decision science or decision-aid science? Eur. J. Oper. Res. **66**, 184–203 (1993)
Rubin, D.B.: Estimating causal effects of treatment s in randomized and non-randomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
Rubin, D.B.: Direct and indirect causal effects via potential outcomes. Scand. J. Stat. **31**, 161–170 (2004)
Savage, L.J.: The theory of statistical decision. J. Am. Stat. Assoc. **46**, 55–67 (1951)
Savage, L.J.: The Foundations of Statistics. Wiley, New York (1954)
Skyrms, B.: Causal Necessity. Yale University Press, New Haven (1979)
Smith, C.A.B.: Consistency in statistical inference and decision. J. R. Stat. Soc. Ser. B Stat. Methodol. **25**, 1–37 (1961)
Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search, Springer, 2nd edn. MIT Press, Cambridge (2000)
Tsoukiàs, A.: On the concept of decision aiding process: an operational perspective. Ann. Oper. Res. **154**, 3–27 (2007)
Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. J. Risk. Uncertain. **5**, 297–323 (1992)
Von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior, 2nd edn. Princeton University Press, Princeton (1947)

Woodword, J.: Making things happen: a theory of causal explanation. In: Oxford Studies in the Philosophy of Science. Oxford University Press, New York (2003)
White, H., Chalak, K.: A Unified Framework for Defining and Identifying causal Effects, UCSD Department of Economics Working Paper, pp. 1–53 (2006)
Wright, S.: Correlation and causation. J. Agric. Res. **20**, 557–585 (1921)

Contributions of Italian Statisticians to the Development of Multivariate Data Analysis

Renato Coppi and Paolo Giordani

Abstract

The main contributions of Italian statisticians to the methodology of multivariate data analysis are investigated, focusing specifically on the development of techniques for coping with the extraction of information from complex data characterized by two or more variables or sets of variables as observed on one or more sets of objects. In particular the following types of methodological areas are considered: supervised and unsupervised classification, regression, factorial and scaling approaches. Although the bulk of this study is devoted to the works appeared in the last three or four decades, some hints are given to the historical profile of the Italian school of Statistics. In this connection it is underlined that the more recent developments are characterized by specific traits of originality, which place the Italian contributions to the aforementioned fields of research somehow at the crossroads among the French, the Dutch, and the Anglo-American schools of Statistics.

1 Introduction

Data analysis has historically been characterized by a "dualistic" perspective. On one side, statistical data have been looked on as being generated by a probabilistic model most often expressed in parametric terms. Typically, inferential procedures based on the likelihood have been adopted for analyzing the empirical information conveyed by the data. The Anglo-American school of thought has greatly influenced this approach. On the other side, the data have been viewed as containing an intrinsic information, independently of any prior knowledge concerning their generation

R. Coppi (✉) • P. Giordani
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
e-mail: renato.coppi@uniroma1.it; paolo.giordani@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7__2,

process. In this case, the analytical procedures have been devised in such a way as to discover and display the empirical information by means of appropriate representations of the data in geometric or algebraic structures, utilizing the formal properties of these structures for carrying out the analyses [Principal Component Analysis (PCA), Correspondence Analysis, Cluster Analysis are examples of this approach, if we look at them from an "exploratory" viewpoint]. The French school of "Analyse des Données" has typically represented this line of research particularly in the period 1960–1980. A third perspective has appeared in the literature in the 1980s and 1990s, mainly inspired by the Dutch school of data analysis (Gifi 1990). This may be defined as "anti-dogmatic," since it is based on flexibility and eclecticism. Although the observed data must drive the procedures of information extraction, an instrumental role is also assigned to statistical models including probabilistic mechanisms that enable the investigator to validate and generalize the results obtained from the analysis of the available data. Alternating least squares algorithms constitute the basic computational tools for implementing the analytical procedures; resampling techniques are often adopted for testing and validation. The school of "Statistical Learning" (Vapnik 1986; Hastie et al. 2009) has many points in common with the latter line of thought, although it puts more emphasis on the probabilistic tools due to the fact that an underlying unknown stochastic process of data generation is assumed and the predictive perspective of the analysis is strongly underlined.

In our paper, we focus on the contributions of Italian statisticians to the domain of multivariate data analysis, restricting our attention to the developments of the "exploratory" approach, including the possible use of probabilistic tools (an interesting previous review of the Italian contributions to this field of analysis is by Balbi 1994). In this respect, it can be stated that the lines of research of Italian scholars have taken inspiration from each of the abovementioned schools of thought, thus realizing a sort of compromise among different ways of looking at the data, with the aim of enhancing and representing in various manners the information embodied in the empirical observations. In this connection, it is evident the link with the traditional characteristics of the Italian school of Descriptive Statistics. One basic feature of this school has historically been the endeavor to incorporate in the statistical tools for analyzing real-world phenomena their *complexity*, without introducing too many theoretical assumptions. While in the first half of the twentieth century, this attitude has mainly produced a wide range of descriptive statistical indices for both univariate and multivariate set-ups, in the second part of the century and more remarkably in the last three decades it turned towards the construction of methods for handling the information contained in complex dependence and interdependence structures among and between sets of variables and for detecting meaningful typologies of statistical units. At the same time, a specific interest has been focused on the analysis of what can be called "complex" statistical observations. In this category we include multiway data arrays as well as special empirical objects like interval-valued or fuzzy-valued variables or, more generally, symbolic data.

In the sequel we will mention some of the numerous contributions provided by Italian statisticians to the above fields of methodological research. In doing that we will limit ourselves to the works whose inspiration is more close to the exploratory perspective rather than to the inferential probabilistic viewpoint. Of course this does not exclude that some of the considered contributions are, in some sense, "model-based" and make use of probabilistic tools. As a matter of fact, this use might be interpreted in the abovementioned framework of the Dutch school, i.e. just as technical means for drawing information from the data and possibly evaluating its statistical reliability. Another preliminary remark concerns the selection of works that we will consider. Due to the practical impossibility of covering the great deal of interesting contributions in the fields under investigation, we have limited the illustration to a small part of the scientific production of interest, trying to enhance the main lines of research rather than detailing the specific proposals. Therefore, we apologize in advance for all the citations that have been missed, while we recognize the value of the overall contribution of Italian scholars to the domain of multivariate data analysis, with particular reference to the many statisticians involved in the activity of the CLADAG section of the Italian Statistical Society. The following discussion is divided into four parts, referring, respectively, to: Regression and Classification, Association structures, Interval and Fuzzy Data, Multiway Data. Due to the limited size of the present paper, the illustration will be very schematic, thus the reader is invited to look at the bibliographic references for getting a deeper insight into the various topics.

2 Regression and Classification

Several lines of research have been followed in this domain, with the common aim of taking into account the complexity of real-world dependence structures for both quantitative and qualitative response variables. In this perspective, some of the works look for improving methods and techniques of common use, while some other works suggest new approaches or propose new methods of analysis.

Noticeable contributions to classification and regression procedures based on decision trees have been provided by Conversano, Cappelli, Mola, and Siciliano (see, e.g., Cappelli et al. 2002; Conversano and Siciliano 2009; Siciliano and Mola 2000). These contributions concern several aspects, including the improvements of decision trees methodology and its use in various data analysis situations (data mining, imputation of missing data, etc.). In this area several other authors provided interesting proposals. For instance, Miglio and Soffritti (2005) introduced suitable proximity measures between Classification trees, which can also be utilized for finding parsimonious optimal trees from among a set of possible trees (Miglio and Soffritti 2004). Another direction of research consists in enlarging the scope of the classical CART trees to include, for example, ordinal response variables. Suitable splitting criteria have been introduced by Piccarreta (2004) to this aim, utilizing a new measure of association between categorical and ordinal variables (Piccarreta 2001).

Concerning traditional algorithms of cluster analysis, such as the k-means method, various authors have provided useful suggestions for improving their efficiency in detecting underlying typologies which can be captured by looking at the empirical features shown by the distribution of the observed units in the space. The use of appropriate metrics and of suitable weights for the observations allows for remarkable improvements in this direction (see, e.g., Cerioli 2001; Cerioli and Zani 2001). The problem of a simultaneous dimensionality reduction and clustering procedure has been faced by Vichi, Rocci, and Vicari in several works, with reference to units $\times$ variables matrices (see, e.g., Rocci et al. 2011; Timmerman et al. 2010; Vichi and Kiers 2001) and also with respect to proximity matrices by means of an appropriate utilization of multidimensional scaling (Kiers et al. 2005). The problem of clustering variables rather than units has been, for instance, faced by Laghi and Soffritti (2005) who proposed a procedure based on suitable measures of collinearity within groups (clusters). Along this line of research a generalized double k-means technique has also been proposed by Vichi and Rocci for simultaneously clustering the units on one side and the variables on the other side (Vichi and Rocci 2008; Vichi and Saporta 2009). In the field of model-based clustering, Galimberti and Soffritti (2007) introduced model-based procedures for detecting multiple cluster structures, namely typologies of units based on different subsets of a set of observed variables.

A new perspective for regression and classification studies was introduced by the Statistical Learning approach, which puts particular emphasis on the predictive viewpoint and on the use of computer intensive techniques based on boosting, bagging and random forests procedures, enabling the researcher to improve the predictive efficiency of regression and classification models. Systematic contributions to this line of research have been provided by Di Ciaccio and Borra in several works. For instance, in Borra and Di Ciaccio (2002) they use bagging and boosting for improving the prediction capability of non-parametric regression techniques. In Borra and Di Ciaccio (2010) they investigate various ways of measuring the prediction error in regression and classification problems and propose some methodological improvements. In the same line, other authors, like Sandri and Zuccolotto (2006), suggest how to use random forests for selecting predictors in a classification problem. Giordano et al. (2004) devise suitable resampling procedures for selecting variables in neural network models for regression analysis and, more generally, contribute to the regression and classification methods based on neural networks (e.g., Perna and Giordano 2001).

3 Association Structures

This is a wide field of analysis whose objective is to detect and display observable or latent structures of association within and between sets of variables. In the following we will just mention some interesting lines of research to which the Italian statisticians gave a noticeable contribution.

Starting from the seminal paper by D'Ambra and Lauro (1982) concerning the analysis of multivariate data with respect to reference subspaces, a line of research has been developed in the field of association structures, consisting in an asymmetric approach to the analysis of relationships among several variables taking into account the influence of "external" variables. The paper by Lauro and D'Ambra (1992) well illustrates this line of thought, while in the works by Beh and D'Ambra (2009); Lombardo et al. (2007, 1996) it is shown how the application of this approach to three-way contingency tables gives rise to a new method called nonsymmetric correspondence analysis, allowing to deal with ordinal variables. In the last decades, different authors tried to decompose association measures for three-way contingency tables. In this perspective, papers by Italian scholars involved in the analysis of three-way contingency tables can be found in Beh et al. (2007); Lombardo (2011); Lombardo et al. (1996); Siciliano and Mooijaart (1997); Simonetti et al. (2010).

Another line of research refers to the improvement and utilization of the Partial Least Squares (PLS) approach in the analysis of several types of association structures. Esposito Vinzi and various co-authors introduce the "Generalized PLS Regression" model (Bastien et al. 2005), which basically allows a transformation of the explanatory variables into orthogonal components in order to improve the fitting to a quantitative or qualitative response variable, by using an appropriate iterative non-linear least squares algorithm. Esposito Vinzi, Lauro and others (Tenenhaus et al. 2005) discuss and improve the procedures of "PLS path modelling," which adopts the PLS approach for estimating suitable structural equation models. Moreover, the authors show that it can provide a general framework for analyzing a multi-block structure of observed variables. A special issue of Computational Statistics and Data Analysis (Esposito Vinzi and Lauro 2005), edited by Esposito Vinzi and Lauro, has been devoted to the various methodological improvements of the PLS approach in several areas of multivariate analysis.

Multidimensional Scaling has been also studied by Italian statisticians. Combining the information contained in the proximities between statistical units with the values taken by "external" variables on those units may lead to a deeper analysis of association through the use of multidimensional scaling procedures which allow also useful visualizations of the results. This line of research is witnessed, for instance, by the works of Bove and Rocci with particular reference to asymmetric proximities (see, e.g., Rocci and Bove 2002).

4 Interval and Fuzzy Data

Interval data, fuzzy data and, more generally, symbolic data represent complex observations requiring particular techniques of analysis. This "complexity" may be due to: (1) imprecision/vagueness of the observed variables, (2) intrinsic complexity of the observed phenomena (represented as "symbolic objects"). Source (1) constitutes a type of uncertainty which needs special methods of treatment, (2) involves a specific mathematical representation of "non-standard" variables. We focus our

illustration on type (1) data, which represent anyway relevant instances of symbolic data (to whose study are addressed many works, in particular, by Verde; see, e.g., Irpino and Verde 2008; Lauro et al. 1998).

Italian statisticians have given remarkable contributions to the analysis of interval data, generalizing in various ways the techniques of PCA, Cluster Analysis and Regression, in order to make them suitable for the analysis of this type of complex data. Concerning PCA, Lauro and Palumbo (2000, 2003) and subsequently Gioia and Lauro (2005) have worked on intervals represented as boxes (or hyperrectangles) in Euclidean spaces looking at their vertices and ranges. Coppi et al. (2003) and D'Urso and Giordani (2004) utilize a "data reconstruction" approach based on a Midpoint-Radius representation of the intervals. Approaches to cluster analysis of interval data have been proposed by, for instance, D'Urso and Giordani (2006) and Irpino and Tontodonato (2006).

Also in the field of fuzzy data many important contributions of the Italian statisticians are to be recorded. A systematic program of re-interpretation of the classical PCA, clustering and regression techniques in terms of fuzzy data is being realized. In particular, Coppi, D'Urso, Ferraro, and Giordani have produced a series of papers in this direction. Basic ingredients of these extensions are: (1) the formalization of fuzzy observations in terms of LR fuzzy numbers; (2) the introduction of appropriate metrics for LR fuzzy variables; (3) the construction of specific models for each of the above fields of analysis (PCA, cluster analysis, regression), incorporating the fuzziness of the observations; (4) the definition of suitable algorithms for estimating the parameters of the introduced models (generally iterative least squares procedures); (5) the possible utilization (in particular for regression analysis) of the notion of Fuzzy Random Variable, which enables the investigator to cope simultaneously with the uncertainty stemming from the imprecision of the data, and the one due to an assumed probabilistic mechanism generating the observations. A partial list of the contributions provided in the above framework is as follows. Concerning PCA of fuzzy data (Coppi et al. 2006b; Giordani and Kiers 2004). For cluster analysis of fuzzy data and fuzzy multivariate trajectories (Coppi et al. 2012; D'Urso and Giordani 2006). As to regression analysis with fuzzy response and crisp or fuzzy explanatory variables (Coppi et al. 2006; D'Urso 2003; Ferraro et al. 2010). Finally, it must be underlined the publication of a special issue of Computational Statistics and Data Analysis, edited in 2006 by Coppi et al. (2006a), devoted to the recent developments of Fuzzy Statistical Analysis.

5 Multiway Analysis

Data generally refer to the observations of some variables on a set of units and are stored in a (two-way) matrix. However, in several situations, data on a set of units on some variables are assumed to be collected in different occasions, leading to a (three-way) array.

In this section, we shall limit our attention to the Italian contributions to the topic of multiway data analysis distinguishing two lines of research, namely component models and cluster analysis.

With respect to component models, at the beginning, the Italian statisticians involved in multiway analysis took inspiration from the French school. In this connection we cite the works by Bolasco (1986), Coppi (1986), and Coppi and Zannella (1978). Notice that in Coppi and Zannella (1978) the so-called Dynamic Factor Analysis has been introduced. Later, it has been extended and generalized (see Coppi and D'Urso 2002; Coppi et al. 1999; Corazziari 1999). In the following years, the Italian statisticians spread their research interests to the methods inspired also from the other schools acquiring an impressive importance within the world multiway community. The success of the meeting "Multiway'88" held in Rome bears witness to the Italian contributions to multiway analysis. The proceedings of the conference (Coppi and Bolasco 1989) represent a milestone within the domain and witness the active role played by Italian researchers. The most relevant findings of Italian statisticians to the Tucker3 and Candecomp/Parafac models and related methods were given by Rocci (1992), Rocci and Giordani (2010), Rocci and Ten Berge (2002). Extensions of such models to imprecise and/or vague data were proposed in Giordani (2010), Giordani and Kiers (2004). Other contributions on component methods can be found in Rizzi and Vichi (1995a,b). Finally, it is useful to mention the special issue of Computational Statistics and Data Analysis on multiway models edited by Coppi and Di Ciaccio (1994), which offered the current (in 1994) world state-of-art of multiway analysis.

The clustering problem for multiway data has been deeply addressed by Italian statisticians. As far as we know, Rizzi (1989) was the first one involved. At least three lines of research can be highlighted. The first one concerns the attempt to look for a sort of compromise partition built on the basis of the set of partitions for all the occasions. Therefore, the research interest is to find the single partition synthesizing at best, according to a given criterion, K available two-way partitions (one for each occasion). In this connection the contributions of Vichi deserve to be cited (Gordon and Vichi 2001; Vichi 1998). Another approach consists of seeking clusters of units considering the data array as a whole, in the sense that no distinct partitions for each occasion are assumed. Thus, in this respect, the data taxonomy is sought by considering the features of the units as the information on a number of variables in different occasions. Rocci, Vicari, and Vichi played an active role in this domain (Rocci and Vichi 2005; Vicari and Vichi 2009; Vichi et al. 2007). Notice that a few of these papers also involved a reduction of the entities of the variable and occasion modes through a Tucker3 model. Finally, the third line of research is about fuzzy clustering for time trajectories. In this case, the occasion mode is the time and data consist of the same variables observed on a number of units in different time occasions. Every unit can then be seen as a multivariate time trajectory. The aim of the analysis is to find a limited number of clusters composed by trajectories homogeneous according to suitable dissimilarity measures. Findings in this domain are due mainly to D'Urso (2000, 2004) also in collaboration with Coppi and D'Urso (2003), Coppi et al. (2010).

References

Balbi, S.: L'Analisi Multidimensional del Dati negli Anni Novanta. Rocco Curto Editore, Napoli (1994)

Bastien, P., Esposito Vinzi, V., Tenenhaus, M.: PLS generalized regression. Comput. Stat. Data Anal. **48**, 17–46 (2005)

Beh, E., D'Ambra, L.: Some interpretative tools for non-symmetrical correspondence analysis. J. Classif. **27**, 55–76 (2009)

Beh, E., Simonetti,B., D'Ambra, L.: Partitioning a non-symmetric measure of association for three-way contingency tables. J. Multivar. Anal. **98**, 1391–1411 (2007)

Bolasco, S.: Per una teoria sulla costruzione e l'analisi delle matrici a tre modi. In: Proceedings of the 33rd Meeting of the Italian Statistical Society, pp. 183–195 (1986)

Borra, S., Di Ciaccio, A.: Improving nonparametric regression methods by bagging and boosting. Comput. Stat. Data Anal. **38**, 407–420 (2002)

Borra, S., Di Ciaccio, A.: Measuring the prediction errors: a comparison of cross-validation, bootstrap and covariance penality methods. Comput. Stat. Data Anal. **54**, 2976–2989 (2010)

Cappelli, C., Mola, F., Siciliano, R.: A statistical approach to growing a reliable honest tree. Comput. Stat. Data Anal. **38**, 285–299 (2002)

Cerioli, A.: Elliptical clusters and the K-means algorithm. In: Book Short Papers CLADAG, Istituto di Statistica, Università degli Studi di Palermo, Palermo, pp. 13–16 (2001)

Cerioli, A., Zani, S.: Exploratory methods for detecting high denosty regions in clusters analysis. In: Borra, S., Rocci, R., Vichi, M., Schader, M. (eds.) Advances in Classification and Data Analysis, pp. 11–18. Springer, Berlin (2001)

Conversano, C., Siciliano, R.: Incremental tree-based missing data imputation with lexicographic ordering. J. Classif. **26**, 361–379 (2009)

Coppi, R.: Analysis of three-way data matrices based on pairwise relation measures. In: De Antoni, F., Lauro, C.N., Rizzi, A. (eds.) Proceedings in Computational Statistics, pp. 129–139. Physica-Verlag, Wien (1986)

Coppi, R., Bolasco, S. (eds.): Multiway Data Analysis. North Holland, Amsterdam (1989)

Coppi, R., Di Ciaccio, A. (eds.): Multiway data analysis: software and applications (Special issue). Comput. Stat. Data Anal. **18**(1), 3–184 (1994)

Coppi R., D'Urso, P.: The dual dynamic factor analysis model. In: Gaul, W., Ritter, G. (eds.) Classification, Automation, and New Media, pp. 47–55. Springer, Heidelberg (2002)

Coppi, R., D'Urso, P.: Three-way fuzzy clustering models for LR fuzzy time trajectories, Comput. Stat. Data Anal. **43**, 149–177 (2003)

Coppi, R., Zannella, F.: L'analisi fattoriale di una serie temporale multipla relativa allo stesso insieme di unità statistiche. In: Proceedings of the 29th Meeting of the Italian Statistical Society, pp. 61–77 (1978)

Coppi, R., Blanco, J., Camaño, G., Corazziari, I.: Descomposición factorial y regresiva de la variabilidad de un array a tres vias. Quantum **4**, 81–107 (1999)

Coppi, R., D'Urso, P., Giordani, P.: Data reduction models for interval valued observations. In: Book of Short Papers CLADAG 2003 (CLADAG'2003, Bologna), pp. 119–122 (2003)

Coppi, R., D'Urso, P., Giordani, P., Santoro, A.: Least squares estimation of a linear regression model with LR fuzzy response. Comput. Stat. Data Anal. **51**, 267–286 (2006)

Coppi, R., Gil, M.A., Kiers, H.A.L. (eds.): The fuzzy approach to statistical analysis (Special issue). Comput. Stat. Data Anal. **51**(1), 1–452 (2006)

Coppi, R., Giordani, P., D'Urso, P.: Component models for fuzzy data. Psychometrika **71**, 733–761 (2006)

Coppi, R., D'Urso, P., Giordani, P.: A fuzzy clustering model for multivariate spatial time series. J. Classif. **27**, 54–88 (2010)

Coppi, R., D'Urso, P., Giordani, P.: Fuzzy and possibilistic clustering for fuzzy data. Comput. Stat. Data Anal. **56**, 915–927 (2012)

Corazziari, I.: Dynamic Factor Analysis. In: Vichi, M., Opitz, O. (eds.) Classification and Data Analysis, Theory and Application, pp. 171–178. Springer, Heidelberg (1999)
D'Ambra, L., Lauro, N.C.: Analisi in componenti principali in rapporto ad un sottospazio di riferimento. Riv. Stat. Appl. **15**, 51–67 (1982)
D'Urso, P.: Dissimilarity measures for time trajectories. J. Ital. Stat. Soc. **1–3**, 53–83 (2000)
D'Urso, P.: Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput. Stat. Data Anal. **42**, 47–72 (2003)
D'Urso, P.: Fuzzy c-means clustering models for multivariate time-varying data: Different approaches. Int. J. Uncertain. Fuzz. Knowl. Syst. **12**, 287–326 (2004)
D'Urso, P., Giordani, P.: A least squares approach to principal component analysis for interval valued data. Chemometr. Intell. Lab. Syst. **70**, 179–192 (2004)
D'Urso, P., Giordani, P.: A weighted fuzzy c-means clustering model for symmetric fuzzy data. Comput. Stat. Data Anal. **50**, 1496–1523 (2006)
D'Urso, P., Giordani, P.: A robust fuzzy k-means clustering model for interval valued data. Comput. Stat. **21**, 251–269 (2006)
Esposito Vinzi, V., Lauro, N.C. (eds.): Partial least squares (Special issue). Comput. Stat. Data Anal. **48**(1), 1–220 (2005)
Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A.: A linear regression model for imprecise response. Int. J. Approx. Reason. **51**, 759–770 (2010)
Galimberti, G., Soffritti, G.: Model-based methods to identify multiple cluster structures in a data set. Comput. Stat. Data Anal. **52**, 520–536 (2007)
Gifi, A.: Nonlinear Multivariate Data Analysis. Wiley, New York (1990)
Gioia, F, Lauro, N.C.: Principal component analysis with interval data. Comput. Stat. **21**, 343–363 (2005)
Giordani, P.: Three-way analysis of imprecise data. J. Multivar. Anal. **101**, 568–582 (2010)
Giordani, P., Kiers, H.A.L.: Principal component analysis of symmetric fuzzy data. Comput. Stat. Data Anal. **45**, 519–548 (2004)
Giordani, P., Kiers, H.A.L.: Three-way component analysis of interval valued data. J. Chemometr. **18**, 253–264 (2004)
Giordano, F., La Rocca, M., Perna C.: Bootstrap variable selection in neural networks regression models. In: Bock, H.-H., Chiodi, M., Mineo A. (eds.) Advances in Multivariate Data Analysis, pp. 109–120. Springer, Heidelberg (2004)
Gordon, A.D., Vichi, M.: Fuzzy partition models for fitting a set of partitions. Psychometrika **66**, 229–247 (2001)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009)
Irpino, A., Tontodonato, V.: Clustering reduced interval data using Hausdorff distance. Comput. Stat. **21**, 271–288 (2006)
Irpino, A., Verde, R.: Dynamic clustering of interval data using a Wasserstein-based distance. Pattern Recognit. Lett. **29**, 1648–1658 (2008)
Kiers, H.A.L., Vicari, D., Vichi, M.: Simultaneous classification and multidimensional scaling with external information. Psychometrika **70**, 433–460 (2005)
Lauro, N.C., D'Ambra, L.: Non-symmetrical exploratory data analysis. Stat. Appl. **4**, 511–529 (1992)
Lauro, N.C., Palumbo, F.: Principal component analysis of interval data: a symbolic data anlysis approach. Comput. Stat. **15**, 73–87 (2000)
Lauro, N.C., Palumbo, F.: Some results and new perspectives in principal component analysis for interval data. In: Book of Short Papers CLADAG, pp. 237–244 (2003)
Lauro, N.C., Giordano, G., Verde, R.: A multidimensional approach to conjoint analysis. Appl. Stoch. Model. Data Anal. **14**, 265–274 (1998)
Laghi, A., Soffritti, G.: A collinearity based hierarchical method to identify clusters of variables. In: Vichi, M., Monari, P., Mignani, S., Montanari, A. (eds.) New Developments in Classification and Data Analysis, pp. 55–62. Springer, Berlin (2005)

Lombardo, R.: Three-way association measure decompositions: the delta index. J. Stat. Plan. Inference **141**, 1789–1799 (2011)
Lombardo, R., Beh, E., D'Ambra, L.: Non-symmetrical correspondence analysis with ordinal variables using orthogonal polymonials. Comput. Stat. Data Anal. **52**, 566–577 (2007)
Lombardo, R., Carlier, A., D'Ambra, L.: Nonsymmetric correspondence analysis for three-way contingency tables. Methodologica **4**, 59–80 (1996)
Miglio, R., Soffritti, G.: Proximity measures between classification trees. In: Bock, H.-H., Chiodi, M., Mineo, A. (eds.) Advances in Multivariate Data Analysis, pp. 27–37. Springer, Heidelberg (2004)
Miglio, R., Soffritti, G.: Simplifying classification trees through consensus methods. In: Vichi, M., Monari, P., Mignani, S., Montanari, A. (eds.) New Developments in Classification and Data Analysis, pp. 31–37. Springer, Berlin (2005)
Perna, C., Giordano, F.: The hidden layer size on feed-forward neural networks: a statistical point of view. Metron Int. J. Stat. **59**, 217–227 (2001)
Piccarreta, R.: A new splitting criterion for classification trees in the ordinal case. In: Bock, H.-H., Chiodi, M., Mineo A. (eds.) Advances in Multivariate Data Analysis, pp. 39–51. Springer, Heidelberg (2004)
Piccarreta, R.: A new measure of nominal-ordinal association. J. Appl. Stat. **28**, 107–120 (2001)
Rizzi, A.: Clustering per le matrici a tre vie. Statistica **49**, 195–208 (1989)
Rizzi, A., Vichi, M.: Representation, synthesis, variability and data preprocessing of a three-way data set. Comput. Stat. Data Anal. **19**, 203–222 (1995)
Rizzi, A., Vichi, M.: Three-way data set analysis. In: Rizzi, A. (ed.) Some Relations Between Matrices and Structures of Multidimensional Data Analysis, pp. 93–166. Giardini Editori e Stampatori, Pisa (1995)
Rocci, R.: Three-mode factor analysis with binary core and orthonormality constraints. J. Ital. Stat. Soc. **1**, 413–422 (1992)
Rocci, R., Bove, G.: Rotation techniques in asymmetric multidimensional scaling. J. Comput. Graph. Stat. **11**, 405–419 (2002)
Rocci, R., Giordani, P.: A weak degeneracy revealing decomposition for the CANDECOMP/PARAFAC model. J. Chemometr. **24**, 57–66 (2010)
Rocci, R., Ten Berge, J.M.F.: Transforming three-way arrays to maximal simplicity. Psychometrika **67**, 351–365 (2002)
Rocci, R., Vichi, M.: Three-mode component analysis with crisp or fuzzy partition of units. Psychometrika **70**, 715–736 (2005)
Rocci, R., Gattone, A., Vichi, M.: A new dimension reduction method: factor discriminant Kmeans. J. Classif. **28**, 210–226 (2011)
Sandri, M. Zuccolotto, P.: Variable selection using random forests. In: Zani, C., Cerioli, A., Riani, M., Vichi, M. (eds.) Data Analysis, Classification and the Forward Search, pp. 263–270. Springer, Heidelberg (2006)
Siciliano, R., Mola, F.: Multivariate data analysis through classification and regression trees. Comput. Stat. Data Anal. **32**, 285–301 (2000)
Siciliano, R., Mooijaart, A.: Three-factor association models for three-way contingency tables. Comput. Stat. Data Anal. **24**, 337–356 (1997)
Simonetti, B., Beh, E.J., D'Ambra, L.: The analysis of dependence for three ways contingency tables with ordinal variables: a case study of patient satisfaction data. J. Appl. Stat. **37**, 91–103 (2010)
Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, N.C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)
Timmerman, M., Ceulemans, E., Kiers, H.A.L., Vichi, M.: Factorial and reduced K-means reconsidered. Comput. Stat. Data Anal. **54**, 1858–1871 (2010)
Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1986)
Vicari, D., Vichi, M.: Structural classification analysis of three-way dissimilarity data. J. Classif. **26**, 121–154 (2009)

Vichi, M.: Principal classifications analysis: a method for generating consensus dendograms and its application to three-way data. Comput. Stat. Data Anal. **27**, 311–331 (1998)
Vichi, M., Kiers, H.A.L.: Factorial k-means analysis for two-way data. Comput. Stat. Data Anal. **37**, 49–64 (2001)
Vichi, M., Rocci, R.: Two-mode multipartitioning. Comput. Stat. Data Anal. **52**, 1984–2003 (2008)
Vichi, M., Saporta, G.: Clustering and disjoint principal component analysis. Comput. Stat. Data Anal. **53**, 3194–3208 (2009)
Vichi, M., Rocci, R., Kiers, H.A.L.: Simultaneous component and clustering models for three-way data: Within and between approaches. J. Classif. **24**, 71–98 (2007)

The Semantic Role of Variability in the Development of Statistical Thought

Paola Monari

Abstract

Since the birth of modern sciences, the development of statistical thought has run along the evolution of the semantic concept of variability. The variability of the natural and social phenomena was the true challenge that Galilean science has faced substituting the order of scientific laws to the apparent disorder of facts. Those laws tried to combine two objectives: the explanation of phenomena in a causal context, and the forecasting of unknown events. In the twentieth century, the most revolutionary scientific theories have been very powerful as explanatory models, but weak as predictive models with reference to single events. All this because the new theories were first of all statistical ones, for example, the theory of evolution for natural selection, the genetics of population, or the quantum physics. Sciences learned to deal with statistic populations and collective properties. The intrinsic characteristics of this kind of laws were properties concerning a phenomenon as a whole, not its inessential micro components.

1 The Principle of Classification: How to Neutralize the Immanence of Variability

Since the birth of modern science the development of statistical thought goes alongside the evolution of the semantic concept of variability. The variability of natural and social phenomena has been the challenge faced by Galilean science by replacing the apparent disorder found in Nature with the order of scientific laws. The need to investigate phenomena producing several different results has shifted the

P. Monari (✉)
Department of Statistical Sciences, University of Bologna, Italy
e-mail: paola.monari@unibo.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_3,

interest of scientific research from the single case to all cases as a whole. The search for laws concerning a group considered as a whole has found its empirical ground in the variability of reality, and of the phenomena that constitute it.

1.1 The Gnosiological Strength of Classification

Scientific knowledge responds to the need for simplification with respect to the multitude of aspects and manifestations in which reality appears to our senses. To classify means grouping together the single items that make up a population, according to similarities and differences with respect to some characteristics, replacing the plurality of individuals with the typology of the classes. Through the principle of classification it is possible to understand the statistical properties of a group only by considering as essential the characteristics according to which similarities and differences are recognized, and by ignoring the many other characteristics that make the single observations appear heterogeneous (Scardovi et al. 1983).

In this attempt to define the classification process followed by modern science, much circularity emerges, which require us to accept a priori some concepts as postulates. We should define what a population is and which are its elementary constituents, what a phenomenon is and what are the relational properties in which it occurs, that is, the latent factors that determine empirical manifestations. Nowadays, modern statistics adopts as a common and shared heritage, many of these concepts, such as "phenomenon," "population," "category," "statistical unit," "elementary event," "characteristic," or "observable variable." Each of these concepts has been investigated by the greatest philosophers, from those of classical antiquity to the greatest statisticians of the twentieth century, who needed to establish their epistemological statements on the new sciences to which they provided their method by renewing it at the roots, beginning with the language.

The ability to classify is innate in human beings and in many animal species. Ordinary language itself has got its basis in classification. Within the common noun is already expressed a classificatory identity that enables us to recognize different entities as the same, only on the basis of a few shared features deemed essential. We find again the "Platonic idea" that associates to each word a class (in itself homogeneous) of facts and things that are similar in some respects (principle of relative similarity), which allows us to recognize what belongs to the class and what is excluded.

1.2 The Phenomenon, a Necessary Abstraction

At this point, however, we cannot neglect a term currently used by scientific language, which, together with the concept of class (category), contains all the strength and the semantic ambiguity of the statistical method. I mean the word *phenomenon*. Karl Pearson writes in *The Grammar of Science* (1892, Chapter II): "... we

have frequently spoken of the classification of facts as the basis of the scientific method, we also have had occasion to use the words real and unreal, and universe phenomenon. It is appropriate, therefore, that before proceeding further we should endeavour to clarify our ideas as to what these terms mean ... But what are these facts in themselves, and what is for us the criterion of their reality" (Pearson 1911). To approach the scientific concept of phenomenon we should revisit the classic Galilean experiment. It did not claim to reproduce reality, but a phenomenon, that is a slice of reality that has been freed from everything that makes it unique and unrepeatable. The same happens when we observe a fragment of reality, even outside an experimental setting.

As the best dictionaries state, "a phenomenon is an observable fact or event, an item of experience or reality, a fact or event in the changing and perceptible forms as distinguished from the permanent essences of things." Well, we must now ask ourselves what is the implied relationship that transforms our perceptions into "facts" related to each others, so as to become inter-subjective and shared macro-concepts. When we define a phenomenon, this loses its historical context and becomes an idealized model that goes beyond contingency; it becomes a conceptual artifact (observational or experimental). Within this idealized model, all circumstances (facts or perceptions) unrelated to those of interest are considered irrelevant. Moreover, the variability induced by circumstances not connected to those considered "strictly related" to the phenomenon is eliminated because it is considered an element of disturbance.

The process of classification extends the epistemological rules of the experimental method to all observable phenomena. The principle of classification is based on a set of relational connections that allow us to isolate what, according to our perceptions, are shared or shareable similarities from what are irrelevant differences, and as such can be virtually eliminated. This process, certainly innate in human beings and consolidated by the need to survive in our natural and social environment, has developed into a rational ability that has led to classify perceptions into homogeneous classes and phenomena, i.e., sets of inter-connected categories. Compared to the concept of class, the concept of phenomenon includes a further abstraction that codifies within a closed system all the relational connections between certain categories of facts, according to a kind of centripetal force, and turns it into a *unicum*, precisely the phenomenon (Krantz et al. 1971–1990).

Modern science has refined these rational abilities and has widely analyzed the philosophical canons that lead from experience to abstract theory. In this context we find the ideal continuity between the statistical and the experimental method. But the path has been very long. Most modern science is based on the concepts of class and phenomenon. Its roots dip into Aristotelian science and pre-Galilean Scholasticisms, which sought its authority in the most extreme classificatory "bulimia." Redemption from those early classification schemes, tarnished by the contamination of ruling esotericisms, was achieved by Linnaeus' *Systema Naturae* (1759), which goes beyond the creationist paradigm that had inspired it and definitely puts the principle of classification amongst the fundamental epistemological canons (Scardovi et al. 1983).

If the statistical method together with the experimental one is the rational foundation of modern sciences, it accompanies the development of many scientific bodies, both in phases of normality, and in those of transformation. On alternate phases, all the sciences have taken advantage from the strategies offered by statistics, and in turn, statistics made use of advantage from the discoveries of other sciences: first astronomy, then biology, physics, psychology, genetics, social sciences, etc. It is a matter of fact that the entire scientific methodology has developed around the many facets in which variability is expressed.

Initially, the aim was to neutralize it to look for invariants. Following the principle of classification, statistics dealt with means, moments, and frequency distributions (Galton 1885). Then, the same principle coherently suggested statistics to identify the different types of variability that observational sciences began to put in evidence. With Pearson (1912) and Fisher (1930b), statistics has built very powerful methods for breaking down variability to compare the (systematic) variability between groups and the (accidental, random) variability within groups. Subsequently, variability was employed to search for the relations that make a phenomenon a conventional concept that can be described by logical or functional relations between its basic components (Galton 1883, 1885; Pearson 1901–1902). We can think about analysis of variance, regression and correlation analysis, exploratory factor analysis, structural equation models, generalized latent variable models, etc.

2 Combinatorial System, Induced Variability, Probability

In the history of the scientific thought, as well as in human history, the concepts of variability and uncertainty have often been associated. They are very different from each other, but also very intertwined, so as to be often confused.

The variability of the physical world has forced man to create coping strategies to find regularities, to make predictions. How is the world beyond our observational perspective, beyond the present time?

It has been more difficult dealing with uncertainty, which is a product of variability, but which concerns the single fact, the individual occurrence. In the distant past as well as today, fortune tellers, oracles, astrologers, magicians have always brought relief to the worries of human beings. Nonetheless, uncertainty has always been a challenge, a mind's creative moment that has manifested in the games of chance (Monari et al. 2005).

One of the first objects intended for this purpose was *astragals* (small sheep or dog's anklebones) invented by man for games of chance, and cited by historians or represented in graffiti, murals, and decorated vases (David 1962). Subsequently, with a decisive leap in abstraction, *astragals* were replaced by dice, artifacts with which man has tried, perhaps unconsciously, to shape a very sophisticated ideal concept, that of symmetry, so to ensure each of the six faces of the die an equal possibility to appear in each throw. The die becomes the symbolic representation of an immutable physical object, which becomes variable when it is used. Even in an

irrational way, man becomes the creator of variability in order to challenge it. That of the dice it is not a phenomenal variability, but pure mental abstraction. One may play dice without dice, just think about all possibilities and pick one.

Therefore, man was familiar with gambling and uncertainty. Why, then, still in ancient times, was not born a mathematics of games that could anticipate modern probability theory in the same way as the forms of the physical world have inspired Euclidean geometry? Many answers have been suggested, all unsatisfactory. We have to take a long time in order to find the first attempts at describing the possible outcomes of games of chance such as throwing dice or coins, attempts that became the empirical premise for the modern combinatorics. However, they still did not mention any measures of potential combinatorial macro-states, seen as aggregations of micro-states (elementary events) that produce the same synthetic result, the outcome (success or failure) of the game.

The history of scientific thought recognizes Luca Pacioli's *Summa de arithmetica*, *geometria*, *proportioni et proportionalità* (1494) and Gerolamo Cardano's *De ludo aleae* (1501–1576), as the forerunners of a new formal language able to describe the space of events in a random experiment, whose dimension is much broader than the few elements that generated the experiment (Hacking 1975). The same language was used by Galilei in his famous essay *Sulla scoperta dei dadi* (1635), where he shows the possible combinations of points in the throw of three dice, whose sum is equal to or less than ten.

Uncertainty could therefore be measured in a rudimentary form consisting of the distribution of all possible events that anticipated the concept of random variable. The awareness of the randomness of events introduces a new rational outlook to the interpretation of the variability of real phenomena, which can be ideally represented by the games of chance, in the same way as the perfect shapes of Euclidean geometry represented physical objects. These real transpositions of abstract concepts (mind experiments or simulations) have offered to a multitude of scholars the intuitive hook to understand the rational foundations of probability and its theorems.

From combinatorics, it is born the idea of a new variability that is no longer the one of the real phenomena, but comes only from the speculative ability of the mind. This new idea of "random variability" is a brilliant and subversive product of the rational thinking that has revolutionized science.

From the work of Cardano almost 100 years had to pass for Pascal (1654) to be able to see in those combinatorial schemes the logical premises of his probability theory. In the language of combinatorics, which is completely deterministic and mathematical, Pascal also found the easiest language to explain to the scientific world the power of his new logic, that of probability. It was a language that did not scare scientists of those times because combinatorial variability remained governed and governable by man, it was a playful mind game that had nothing to do with the reality of phenomena. It remained completely subjugated to that which will become Laplace's determinism, in which probability was confined to neutralize the effects of accidental errors in measurement, in order to search for the "true" value of the observed magnitude.

In that "neutral" context, the first probabilists managed to demonstrate fundamental theorems, just think about De Moivre, Lagrange, Bernoulli, and Gauss (Hacking 1975; Hald 2003; Stigler 1986). These leading figures of modern thinking, however, were not only mathematicians; they were above all physicists and astronomers, and their philosophical speculations were strongly influenced by observational experience. Gauss (1809) drew his famous model in a purely analytical way, after assuming some formal preconditions, which he had taken from the evidence regarding the distribution of the repeated measurements of astronomical magnitudes, following an entirely circular logical path. That evidence had already led Lagrange (1806) to indicate the arithmetic mean of instrumental measurements as the most likely value for an unknown quantity. And he did so, before the adventurous inversion of De Moivre's theorem would have generated the ambiguous confusion with the law of large numbers, logically resolved only with modern statistical inference (Porter 1986).

However, for many years, probability continued to be convenient to compensate for the human mind's cognitive limitations, a mind that could never compete with the "infinite intelligence" postulated by Laplace (1814). But the subtle workings of this new logic and its language were broadening the possible horizons of scientific thinking. Once identified a phenomenon, this could be described by a statistical model able to interpret and manage both the "accidental" variability that differentiates between individuals, and the "systematic" variability that characterizes the phenomenon in its essential trait (Scardovi 1982).

3 The Return Match of Variability

In 1859 Darwin had already published *On the Origin of Species*, and formulated his theory of evolution by natural selection, which offers to science a new way of reading the variability up to then described by deterministic models. In Darwin's theory the species are not immutable but they evolve conditioned by the environment (Darwin 1972). Beyond its ethical and philosophical impact, which has not yet diminished, this theory opened up two huge issues: (1) to prove the new theory in quantitative terms, and (2) to find the processes that determine the phenotypical changes upon which the environment could act selectively.

The first issue promoted the rise of the statistical method with the fundamental works of Galton, Pearson, and Weldon jointly with the journals that have launched statistics in the world as a unifying method of modern sciences: *The Journal of the Royal Statistical Society* and *Biometrika*.

Variability was no longer a state of disorder to be eliminated in order to find the true laws of nature, but became in itself a source of knowledge. The laws of the physical world could be discovered only by studying the variability of its phenomena, namely the set of relationships that conventionally connect the observed facts. In this context, it originates the theory of linear regression in which

is nested the concept of causality, and that of linear correlation, where the concept of cause fades into a state of interdependence. The latter offered to Pearson (1912) and Spearman (1904) the idea of a latent explanation underlying observed phenomena, the "factors," in a constant pursuit of a deep causal system.

The second issue relates more closely to scientific research, and particularly biology in its new structure, that is genetics. Mendel (1866) had the brilliant idea of the genetic inheritance of characters, expressed in the simplest form of a diallelic gene by the expansion of Newton's binomial formula $[p(A) + p(a)]^n$, where the exponent indicates the generations, A and a refer to the dominant and recessive alleles that determine the phenotype, while the binomial coefficients define the numerical proportions of genotypes and phenotypes.

The theorem's structure at the basis of the representation of the hereditary process experimentally demonstrated by Mendel, and by those who came after him to codify molecular genetics and population genetics, is the basic one of the repeated toss of the coin, which describes the aggregation of micro-states (the possible combinatorial outcomes) in macro-states (all the combinations that produce the same expected result) (Monari et al. 2009; Monari and Scardovi 1989). In this new paradigm combinatorics and variability become *modus intellegendi* of a new science that finds in the ancient gambling the most appropriate language to provide semantic, and at the same time formal content to the explanation of its processes.

If analogy has a place in the evolution of scientific thought, then we can understand the growth process of Ronald Fisher, who, starting from the discovery of the life sciences, was led to his extraordinary contributions to statistics (Monari et al. 2009). Fisher gave an original theoretical layout to population genetics (Fisher 1930a) by blending the Darwinian theory of evolution and the Mendelian genetics.

From this huge work what has modern statistics taken, beyond the strength of the method? The answer is: a new way of dealing with variability. The arithmetic mean is no longer the final point of a science that seeks above all the invariants. As a model of invariance, the arithmetic mean becomes the starting point to investigate variability. Standard deviation is no longer the worrying measurement of dispersion or the reassuring measurement of precision. The analysis of variance breaks new grounds because it allows recognizing the variability within groups as a sign of a system in equilibrium, from the variability between groups, as sign of significant differences between groups. Fisher ingeniously associated the first type of variability to that of combinatorial schemes, generated by constant probabilities, and translated it into the language of random sampling, where variability is just sampling error. On the other hand, he associated the second type of variability to the one that occurs when an innovative factor breaks the balance and changes the original connections (parameters, probabilities, etc.). This entirely new perspective of variability has completely transformed modern statistics, which became much more than a mere tool for quantitative research; it became the explanatory language of the new sciences (Monari et al. 2009).

4 Time and Variability

Modern science had to wait for the twentieth century to acknowledge that time is an intrinsic factor in the variability of phenomena. There is the variability expressed by the uncertainty of a future event since the conditions that regulate the phenomenon are unknown: here time is inert, and uncertainty about the future is the same as in the unknown past. And there is the variability that instead is created and shaped by time: time becomes a factor of variability because it intervenes to provide a direction to the phenomena that evolve, in the same way as evolutionary turning points mark the time. When this variability intervenes, time becomes irreversible and phenomena cannot return to their previous state (in the sense that the probability of this occurrence is 0).

The dominant Laplacian philosophy strengthened the thesis that the order of the universe was fixed at the origin, and could remain unchanged and unchangeable in astronomical phenomena as in life phenomena (Spearman 1904). Laplace's thinking has influenced all of modern science, which had strenuously tried to fight the first signs of weakness when, in the second half of the nineteenth century, Charles Darwin disrupted all research canons through a dynamic and evolutionary explanation of natural variation in which time is beaten by the clock of the generations that pass. Darwin's concept of time is a time measured along the direction imprinted by environmental factors on the combinatorial variability of genetic crosses in the passing of generations. For the first time, a time that does not allow return is established. In the same way in which time in Boltzmann's physics did not allow any logical return (Boltzmann 1905).

The question of prediction was then open. In an entirely deterministic world, all events could be predictable. If they are not so, it is just because we do not have the "infinite intelligence" that would allows us to know at any moment all the forces by which nature is moved. For centuries this has allowed humans to foresee large astronomical phenomena, and to classify a living being into its species. Here uncertainty is only a limit of the researcher's skills, which does not remove semantic value from scientific law, and the variability of single components is only a factor of disturbance. Prediction of each single event becomes possible only through cognitive approximation, but it could be exact.

The new science of Darwin and Boltzmann is not like this anymore, it shows another world that is indeterministic and that can only be explained by the language of probability. The living species are no longer immutable, but become a continuous interlaced web of genetic combinations, contingent factors, and environmental contexts. Thus, although a single molecule follows the laws of classical physics, a population of molecules follows other rules that are statistical and combinatorial, and that lead that population toward the most probable state of maximum entropy, driven only by random combinations of elementary events.

Once again, the statistics lends its semantic language and acquires new tools: new measurements for variability in terms of entropy, formalization of stochastic processes, and time series analysis that break down a phenomenon that changes over time in all its possible components (Scardovi 1982).

5 Statistical Laws and Revealing Variability

Galileo's experimental rationality pictured a nature that could be described through the language of mathematics (today statistics), in which qualities could be converted into quantities. This was the kind of science used in astronomy, an observational science that could only emulate the canons of the Galilean experiment. Moreover Galileo wanted to establish a way of thinking free from metaphysical prejudice and anchored in experience. Galileo's rationality is that of Kepler, of the great astronomers who came before him, and those who followed him until Newton. The laws of astronomy were looking for regularities within the intricate web of variability in the movements, sizes, space and, above all, measurements. Those laws had to interpret the divine plan, but also had to convince people by means of the accuracy in the predictions of celestial events.

The laws of Galilean science planned to combine two objectives: (a) to explain phenomena in a causal context, and (b) to predict events not yet explained by those laws. The statements of modern science have not always achieved both objectives. However, the most revolutionary scientific theories of the twentieth century were very powerful as explanatory models, but often very weak as forecasting models when applied to single events. This is because the new theories are first of all "statistical" theories. Science has learned to deal with statistical populations and collective properties.

The characteristics of these laws are properties that relate to a phenomenon as a whole and not to its elementary (inessential) components. Scientific interest shifts from single units to groups in search of statistical regularities which become group properties. For example, the sex ratio at birth is a feature of many species, and it does not apply to a single birth; in the same way as the second law of thermodynamics does not refer to a single molecule, but it describes the possible states of a large set (population) of molecules. The genetic theory of heredity too does not allow determining with certainty how each person will be, but it accurately describes the genetic structure of a group.

What are the conditions according to which a statistical property, a regularity, a distribution of frequency observed in a group may free themselves from the group dimension to become parameters of a population, scientific laws, theories? The law of large numbers attempted to answer this. The answer is not a mathematical theorem, nor a scientific discovery; it is the expression of the rational ability of man to find rules in the repetition of its experiences and perceptions. The repetition of experiences and observations in order to search for regularities is a common feature of scientific research, whatever the factual or epistemological context in which it takes place. The distinction between absolute laws and statistical laws, which has so animated the philosophical debate in the last centuries, is solved statistically in identifying phenomena with no variability from phenomena consisting of several different elementary events in which the variability between single units cannot be eliminated.

When the phenomena of interest for science are "statistical," variability becomes the explanation key and acquires its own semantic meaning. A macro phenomenon is statistically stable because it is the result of many irregular micro phenomena. This means going beyond the single phenomenon and looking for a different perspective: the physical sciences shifted their attention to macro phenomena in a more agile and less absorbing way than life sciences and social sciences.

The first epistemological consequence is that every semantic distinction between sample and population blurs when the sample is large enough to bring out the law; in the same way as, faced with a statistical statement, the distinction between validation and confutation blurs (Scardovi 1988, 1999). In the "statistical laws," the analysis of variability becomes the main focus of scientific research, and statistics plays a main role and it is no longer just a tool. Statistical language becomes the language of the new scientific theories, and the statistical method for the study of variability in all its facets offers the interpretive keys for all phenomena, as well as the conceptual tool to follow the evolution of a phenomenon through the transformations of its internal variation or its entropic system.

Acknowledgments The vindication of a "plagiarism." Dedicated to Italo Scardovi. Addressing the issue of statistics' epistemological statements at a conference dedicated to the 150th anniversary of Italian Statistics brings immediately to our mind the fundamental methodological contributions made by Italo Scardovi, one of the most notable representatives of the Italian school.

References

Boltzmann, L.: Populare Schriften. Barth, Leipzig (1905)
Darwin, C.: On the Origin of Species. John Murray, London (1972)
David, F.N.: Games, God and Gambling. Charles Griffin and Company, London (1962)
Fisher, R.A.: The Genetical Theory of Natural Selection. Clarendon, Oxford (1930a)
Fisher, R.A.: Statistical Method for Research Workers. Oliver and Boyd, Edinburgh (1930b)
Galton, F.: Inquiries into Human Faculty and Its Development. Dent &Co, London (1883)
Galton, F.: Regression towards mediocrity in hereditary stature. J. Royal Anthropol. Inst. **XV**, 263 (1885)
Hacking, I.: The Emergence of Probability. Cambridge University Press, Cambridge (1975)
Hald, A.: History of Probability and Statistics Before 1750. Wiley, New Jersey (2003)
Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of Measurements, vol. I. Academic, New York (1971–1990)
Laplace, P.S.: Essai philosophique sur le probabilités. Gauthier-Villar, Paris (1814)
Monari, P.: The concept of measure, from Plato to modern statistics. Art of living or intellectual principle? Statistica **LXV**, 3, 243–255 (2005)
Monari, P., R.A. Fisher: The relevance of the genetical theory of natural selection. Statistica **LXIX**, 129–142 (2009)
Monari, P., Scardovi, I.: I fondamenti statistici dell'equilibrio genetico nelle popolazioni. Ed. Martello, Milano (1989)
Pearson, K.: On the systematic fitting of curves to observations and measurements. Biometrika **1**, 265–303 (1901–1902)
Pearson, K.: The Grammar of Science, 3rd edn. W. Scott, London (1911)
Pearson, K.: On the general theory of influence of selection on correlation and variation. Biometrika **8**, 437–443 (1912)

Porter, T.M.: The Rise of Statistical Thinking 1820–1900. Princeton University Press, Princeton (1986)
Scardovi, I.: The idea of change in the statistical in tution of natural variability. Genus **38**, 3–4, 1–17 (1982)
Scardovi, I.: Statistical induction: probable knowledge or optimal strategy? Epistemologia **VII**(6), 101–120 (1984)
Scardovi, I.: Ambiguous uses of probability. In: Agazzi, E. (ed.) Probability in the Science. Kluwer Academic Publisher, Dordrecht, 51–66 (1988)
Scardovi, I.: Statistical inference and inductive prevision. In: Slottje, D.J. (ed.) Advances in Econometrics, Income Distribution and Scientific Methodology. Essays in Honor of Camilo Dagum. Physica-Verlag, New York, 301–320 (1999)
Spearman, C.: General intelligence, objectively determined and measured. Am. J. Psychol. **15**, 201–293 (1904)
Stigler, S.M.: The History of Statistics. The Measurement of Uncertainty Before 1900. The Belknap Press of Harvard University Press, Cambridge (1986)

The Permutation Testing Approach in the Light of Conditionality and Sufficiency Principles

Fortunato Pesarin

Abstract

In recent years permutation testing methods have increased in number of applications and in solving complex multivariate problems. When available they are essentially of an exact nature in a conditional context, where the conditioning is on the pooled observed data which in general are a set of sufficient statistics in the null hypothesis. The application of the conditionality principle of inference provides this approach with important and useful properties.

1 Introduction

In recent years permutation testing methods have increased both in number of applications and in solving complex multivariate problems. Most of testing problems may also be effectively solved using traditional parametric or rank-based nonparametric (NP) methods, although in relatively mild conditions their permutation counterparts when available are asymptotically as good as the best ones (Hoeffding 1952). Permutation tests (PTs) are essentially of an exact NP nature in a conditional context, where the conditioning is on the pooled observed data which, under randomization of units to treatments, are always a set of sufficient statistics in the null hypothesis. On the one hand, the application of the conditionality principle (CP) of inference provides the PT approach with important and useful properties. On the other, the reference null distribution of most parametric tests, with the exception of rather simple situations, is only known asymptotically. Thus, for most sample sizes of practical interest, the possible lack of efficiency of PTs

F. Pesarin (✉)
Department of Statistical Sciences, University of Padua, Via Cesare Battisti 241, 35121 Padova, Italy
e-mail: fortunato.pesarin@unipd.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_4,

may be compensated by the lack of approximation of parametric counterparts. There are many complex multivariate problems (common in biostatistics, clinical trials, experimental data, pharmacology, psychology, social sciences, etc.) which are difficult, if not impossible, to solve outside the CP and in particular outside the method of nonparametric combination (NPC) of dependent PTs (Pesarin and Salmaso 2010).

Frequently parametric methods reflect a modelling approach and generally require a set of quite stringent assumptions, which are often difficult to justify. Sometimes these assumptions are merely set on an *ad hoc* basis: too often and without any justification researchers assume multivariate normality, random sampling from a target population, homoscedasticity of responses also in the alternative, random effects independent of units, etc. In this way consequent inferences have no real credibility. On the contrary, NP approaches try to keep assumptions at a lower workable level, avoiding those which are difficult to justify. Thus, they are based on more realistic foundations, are intrinsically robust and consequent inferences credible. For instance, PT comparisons of means do not require data homoscedasticity in the alternative, provided that random effects are either negative or positive.

Our point of view, however, is that statisticians should have in their tool-kit of methods both the parametric, including the Bayesian, and the NP, because in their life they surely meet with problems which are difficult, if not impossible, within one approach and others which in turn are difficult, if not impossible, within the other. For some examples as well as for the literature on the subject matter, we refer to Pesarin and Salmaso (2010) and references therein.

Here we discuss main properties of PTs derived by direct application of sufficiency principle (SP) and CP. The outline includes: a discussion of data model which extends that commonly used by parametric approaches; a presentation of SP and CP and their involvement in the PT principle; notation, definitions, and main properties (exactness, similarity, uniform unbiasedness, consistency) of PTs; and the extension of conditional to unconditional inference.

2 The Data Model

Without loss of generality we refer to the two-sample one-dimensional design as a guide. Extensions to one-sample and multi-sample designs are straightforward. The extension to multivariate designs requires the NPC (Pesarin and Salmaso 2010).

Let us assume that a variable X takes values on sample space $\mathscr{X}$, and that associated with $(X, \mathscr{X})$ there is a parent distribution P member of an NP family $\mathscr{P}$. "A family $\mathscr{P}$ of distributions is NP when it is not possible to find a finite-dimensional space Θ (the parameter space) such that there is a one-to-one relationship between Θ and $\mathscr{P}$, in the sense that each member P of $\mathscr{P}$ cannot be identified by only one member θ of Θ, and vice versa." In practice parametric families only contain distributions defined by a well-specified finite set of parameters; whereas families of distributions in which parameters are infinitely

many or are unspecified are NP. Each $P \in \mathscr{P}$ gives the probability measure to events A member of a suitable collection $\mathscr{A}$ of events. Family $\mathscr{P}$ may consist of distributions of real (continuous, discrete, mixed) and/or categorical (nominal, ordered) type of variables. It is assumed that $\mathscr{P}$ admits the existence of a dominating measure $\xi_{\mathscr{P}}$ in which respect the density $f_P(X) = \mathrm{d}P(X)/\mathrm{d}\xi_{\mathscr{P}}$ is defined. The density on every observed sample point $X \in \mathscr{X}$ is assumed satisfying to $f_P(X) > 0$ (we do not distinguish between a variable X and its observed sample points, the context suffices to avoid misunderstandings).

Let $\mathbf{X}_j = \{X_{ji},\ i = 1, \ldots, n_j\} \in \mathscr{X}^{n_j}$ be the independent and identically distributed (IID) sample data of size n_j coming from $P_j \in \mathscr{P}$, $j = 1, 2$. A notation for data sets with independent samples is $\mathbf{X} = \{X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}\} \in \mathscr{X}^n$, whose related model, with clear meaning of the symbols, is $(\mathbf{X}, \mathscr{X}^n, \mathscr{A}^{(n)}, P^{(n)} \in \mathscr{P}^{(n)})$, where $n = n_1 + n_2$, and $P^{(n)} = P_1^{n_1} \cdot P_2^{n_2}$. To denote data sets in the PT context it can be useful referring to the unit-by-unit representation: $\mathbf{X} = \mathbf{X}^{(n)} = \{X(i),\ i = 1, \ldots, n;\ n_1, n_2\}$, where it is intended that first n_1 data in the list belong to first sample and the rest to the second. Indeed, denoting by $\Pi(\mathbf{u})$ the set of permutations of unit labels $\mathbf{u} = (1, \ldots, n)$ and by $\mathbf{u}^* = (u_1^*, \ldots, u_n^*) \in \Pi(\mathbf{u})$ one of its members, $\mathbf{X}^* = \{X^*(i) = X(u_i^*),\ i = 1, \ldots, n;\ n_1, n_2\}$ is the related permutation of $\mathbf{X}$; so that $\mathbf{X}_1^* = \{X_{1i}^* = X(u_i^*), i = 1, \ldots, n_1\}$ and $\mathbf{X}_2^* = \{X_{2i}^* = X(u_i^*),\ i = n_1 + 1, \ldots, n\}$ are the two permuted samples, respectively. Of course, in multivariate problems data vectors associated with units are then permuted.

We discuss testing problems for stochastic dominance alternatives (one-sided) as are generated by treatments with nonnegative random shift effects Δ. In particular, the alternative assumes that two treatments produce effects Δ_1 and Δ_2, and that $\Delta_1 \overset{d}{>} \Delta_2$, where $\overset{d}{>}$ stands for stochastic (or distributional) dominance. Thus, the hypotheses are $H_0 : X_1 \overset{d}{=} X_2 \equiv P_1 = P_2$, and $H_1 : (X_1 + \Delta_1) \overset{d}{>} (X_2 + \Delta_2)$, respectively. Extensions to nonpositive and two-sided alternatives are straightforward. Note that *under H_0 data of two samples are exchangeable*, in accordance with the notion that units are randomized to treatments. Without loss of generality, we assume that effects in H_1 are such that $\Delta_1 = \Delta \overset{d}{>} 0$ and $\Pr\{\Delta_2 = 0\} = 1$. This condition agrees with the notion that an *active treatment* is only assigned to units of first sample and a *placebo* to those of the second. Moreover, Δ can depend on units and on related null deviates X, so that pairs (X_{1i}, Δ_i), $i = 1, \ldots, n_1$, do satisfy $(X_{1i} + \Delta_i) \geq X_{1i}$ with at least one strict inequality. In this situation the induced stochastic dominance $(X_1 + \Delta) \overset{d}{>} X_2 = X$ is compatible with heteroscedasticities in the alternative. Thus, H_0 can also be written as $H_0 : \Delta \overset{d}{=} 0$. Other than measurability, no further distributional assumption on random effects Δ is required. It is required that null deviates X and test statistics $T : \mathscr{X}^n \to \mathscr{R}^1$ are measurable in H_0. To emphasize the roles of sample sizes and effects, we may use the notation $\mathbf{X}^{(n)}(\Delta) = \{X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}\}$ to denote data sets; and so $\mathbf{X}^{(n)}(0)$ denotes data in H_0. It is worth noting that the pooled data $\mathbf{X}^{(n)}(0)$ is always a set of sufficient statistics for P in

H_0. Indeed, since $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}) = 1$, the conditional distribution of $\mathbf{X}$ given $\mathbf{X}$ is independent of P. Furthermore, when P is NP or the number of its parameters is larger than sample size or when it lies outside the regular exponential family, $\mathbf{X}$ is *minimal sufficient*.

PT lie within the conditional method of inference, the conditioning being on the observed data set $\mathbf{X}$. The related conditional reference space is denoted by $\mathscr{X}^n_{/\mathbf{X}}$. Essentially $\mathscr{X}^n_{/\mathbf{X}}$ is the set of points of sample space $\mathscr{X}^n$ which are equivalent to $\mathbf{X}$ in terms of information carried by the associated underlying likelihood. Thus, it contains all points $\mathbf{X}^*$ such that the likelihood ratio $f_P^{(n)}(\mathbf{X})/f_P^{(n)}(\mathbf{X}^*)$ is P-independent, and so it corresponds to the *orbit* of equivalent points associated with $\mathbf{X}$. Given that, under H_0, the density $f_P^{(n)}(\mathbf{X}) = \prod_{ji} f_P(X_{ji})$ is by assumption exchangeable in its arguments, because $f_P^{(n)}(\mathbf{X}) = f_P^{(n)}(\mathbf{X}^*)$ for every permutation $\mathbf{X}^*$ of $\mathbf{X}$, then $\mathscr{X}^n_{/\mathbf{X}}$, or $\mathscr{X}_{/\mathbf{X}}$ by suppressing superscript n, contains all distinct permutations of $\mathbf{X}$. That is $\mathscr{X}_{/\mathbf{X}} = \{\bigcup_{\mathbf{u}^* \in \Pi(\mathbf{u})}[X(u_i^*), i = 1, \ldots, n]\}$. Therefore, since every element $\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ is a set of sufficient statistics for P in H_0, $\mathscr{X}_{/\mathbf{X}}$ is a *sufficient space*. Conditional reference spaces $\mathscr{X}_{/\mathbf{X}}$ are also called *permutation sample spaces*. Moreover, since $\forall A \in \mathscr{A}$ the conditional probability $\Pr(A|\mathbf{X}) = \Pr(A|\mathscr{X}_{/\mathbf{X}})$ in H_0 is P-independent (**P.1** in Sect. 4), the pooled data set $\mathbf{X}$ can be considered as playing the role of ancillary statistics for the problem. And so, when $\mathbf{X}$ is *minimal sufficient* it is also *maximal ancillary* and unique, except for a permutation.

In paired-data designs what is essential is that in H_0 the distribution of X is symmetric with respect to 0 (Pesarin and Salmaso 2010). This condition can be achieved in two main instances: (a) when data are exchangeable within each unit, i.e. when $Y_{1i} \overset{d}{=} Y_{2i}\ \forall i = 1, \ldots, n$, the Ys being paired responses, in which the difference of any two individual observations in $H_0^A : Y_1 \overset{d}{=} Y_2$ is symmetrically distributed around 0, and the set of differences $\mathbf{X} = \{X_i = Y_{1i} - Y_{2i}, i = 1, \ldots, n\}$ is sufficient for P; (b) when Y_{1i} is symmetric around μ_{1i} and Y_{2i} around μ_{2i} without being homoscedastic (and so not exchangeable), then their difference $Y_{1i} - Y_{2i}$ is symmetric around 0 in $H_0^B : (\mu_{1i} - \mu_{2i} = 0, i = 1, \ldots, n)$. In both instances, however, $\mathscr{X}_{/\mathbf{X}} = \{\bigcup_{\mathbf{S}^* \in [-1,+1]^n}[X_i S_i^*, i = 1, \ldots, n]\}$ contains all points obtained by assigning signs $S = (+1,-1)$ to differences in all possible ways. By the way, paired-data designs show that the exchangeability property is sufficient but not necessary for the PT approach.

The fact that random effects Δ may depend on null deviates X can be viewed as an improvement with respect to traditional parametric approaches, though this may imply evident difficulties for estimation and prediction. On the one hand, this leads to assumptions that are much more flexible and closer to reality. There are indeed many real problems in which the assumption of independence of effects on null deviates cannot be justified, as, for instance, when data are obtained by measurement instruments based on nonlinear monotonic transformations φ of underlying deviates Y. Indeed, with clear meaning of the symbols $\Delta\varphi'(Y + c\Delta) = X(\Delta) - \varphi(Y)$, which depends on $Y = \varphi^{-1}(X)$ and Δ. On the other hand, it is noticeable that in

PTs the separate estimate of variance components is not required. Consequently, the modeling may better fit physical requirements, results of analyses are more credible and their interpretation more clear. In addition, it is to be emphasized that in the NP framework, more than on parameters, the inferential interest is on functionals, i.e. on functions of all parameters such as the so-called treatment effect Δ. And so it can be impossible to separate the role of parameters of interest from the nuisance ones since they could be confounded in Δ.

Moreover, when the data set $\mathbf{X}$ is minimal sufficient in H_0, even if the parent likelihood model depends on a finite set of parameters only one of which is of interest, univariate statistics capable of summarizing the contained information do not exist. So no parametric method can claim to be uniformly better than others. Indeed, conditioning on $\mathscr{X}_{/\mathbf{X}}$, i.e. by considering PT counterparts, improves the power behavior of any unbiased test statistic (via Rao-Blackwell). However, to reduce the loss of information associated with using one single statistic, it is possible to find solutions within the so-called multi-aspect methodology and based on the NPC of several dependent PTs, each capable of summarizing information on a specific aspect of interest for the analysis (Pesarin and Salmaso 2010). A procedure which may improve efficiency and interpretability of results. For instance, when of two unbiased partial PTs only one is consistent, their NPC is consistent.

3 Conditionality, Sufficiency, and Permutation Testing Principles

Let us briefly recall the CP and the SP, as are used in parametric inference (Cox and Hinkley 1974). We consider these principles as key guides also for the NP approach and relate them to the PT principle.

The SP states that: "Suppose that we are working with the model $f_X(x, \theta)$ for the random variable X, according to which the data set $\mathbf{X}$ is observed, and also suppose that the statistic S is minimal sufficient for $\theta \in \Theta$. Then, according to the SP, so long as we accept the adequacy of the model, identical conclusions should be drawn from data $\mathbf{X}_1$ and $\mathbf{X}_2$ with the same value of S."

The CP states that: "Suppose that C is an ancillary statistic for the problem, then any conclusion about the parameter or the functional of interest is to be drawn as if C were fixed at its observed value."

Basically, the rationale for adopting these principles in statistical inference considers typical examples as the following: suppose that data $\mathbf{X}$ can be obtained by means of one of two different measuring instruments, I_1 and I_2, and suppose the associated normally distributed models are, respectively, $X_1 \sim \mathscr{N}(_1, \sigma_1)$ and $X_2 \sim \mathscr{N}(_2, \sigma_2)$, with $\sigma_1 \ll \sigma_2$. If it is known which instrument has generated $\mathbf{X}$ it seems unavoidable to condition on the related (ancillary) model in any inference regarding μ, the value of σ being known or unknown. Moreover, in accordance with the SP the statistical estimator of unknown μ should be based on a, possibly minimal complete, sufficient statistic for it. In addition, if the nuisance parameter σ is

unknown, it is wise to stay at least on invariant statistics or on the invariance of null rejection probability (according to the notion of similarity) and so to condition on a possibly minimal sufficient statistic for it. Indeed, by acting outside these principles related inferential conclusions can be biased, misleading and maybe impossible to be correctly interpreted.

Thus, in the general situation it is wise to condition on its minimal sufficient statistic in H_0, i.e. to condition on the pooled observed data $\mathbf{X}$ which is always sufficient for whatever $P \in \mathscr{P}$ and ancillary for the inferential problem. It is to be recognized that in the literature there is general agreement on the SP; whereas the CP, especially when the ancillary statistic C is not unique, gives rise to known questions and so it is somewhat doubtful (Frosini 1991). These doubts, however, do not apply to the PT approach when $\mathbf{X}$ is minimal sufficient and so maximal ancillary and unique.

This kind of conditioning implies referring to the PT principle: "If two experiments, taking values on the same sample space $\mathscr{X}$ with underlying distributions P_1 and P_2 give the same data $\mathbf{X}$, then two inferences conditional on $\mathbf{X}$ and obtained by using the same statistic T must be the same, provided that the exchangeability of the data is satisfied in H_0." Of course, it is intended that in order to obtain reliable inferences there must be a form of stochastic dominance of $(T|H_1)$ with respect to $(T|H_0)$.

On the one hand it should be emphasized that the PT principle works in accordance with both CP and SP since it satisfies both. On the other hand, the related conditional inference can be extended from the set of really observed units to the family of all populations whose associated distributions P satisfy the condition $f_P^{(n)}(\mathbf{X}) > 0$, so as to also include most of the problems in which the sample data are obtained by selection-bias procedures from a target population. However, it should be noted that, due to conditioning on sufficient statistics for all nuisance entities, the extension to a family of distributions is also typical of all parametric conditional inferences in the presence of nuisance parameters (Sect. 5). For instance, this feature is clearly enjoyed by Student's t whose inference can be extended from the observed data set $\mathbf{X}$ to all normal populations which assign positive density to the variance estimate $\hat{\sigma}^2$; thus, its inference can be extended to a family of distributions more than to only the target one.

4 Main Properties of PTs

In this section we briefly outline main terminology, definitions, and general theory of PTs for some one-dimensional problems. Emphasis is again on two-sample one-sided designs in which large values of test statistics $T : \mathscr{X}^n \to \mathscr{R}^1$ are evidence against H_0.

- **P.1.** *Sufficiency of $\mathscr{X}_{/\mathbf{X}}$ for P under H_0 implies that the null conditional probability of every event $A \in \mathscr{A}$, given $\mathscr{X}_{/\mathbf{X}}$, is independent of P; that is, with clear meaning of the symbols,* $\Pr\{\mathbf{X}^* \in A; P|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\mathbf{X}^* \in A|\mathscr{X}_{/\mathbf{X}}\}$.

Thus, the permutation distribution induced by any test statistic $T : \mathscr{X}^n \to \mathscr{R}^1$, namely $F_T(t|\mathscr{X}_{/\mathbf{X}}) = F_T^*(t) = \Pr\{T^* = T(\mathbf{X}^*) \leq t|\mathscr{X}_{/\mathbf{X}}\}$, is P-invariant. Hence, any related conditional inference is distribution-free and NP. Moreover, since for finite sample sizes the number $M = M^{(n)} = \sum_{\mathscr{X}_{/\mathbf{X}}} \mathbb{I}(\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}})$ of points in $\mathscr{X}_{/\mathbf{X}}$ is finite, a relevant consequence of both independence of P and finiteness of M is that in H_0 the permutation probability on every $A \in \mathscr{A}$ is calculated as

$$\Pr\{\mathbf{X}^* \in A|\mathscr{X}_{/\mathbf{X}}\} = \sum_{\mathbf{X}^* \in A} f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* \Big/ \sum_{\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}} f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* = \sum_{\mathscr{X}_{/\mathbf{X}}} \frac{\mathbb{I}(\mathbf{X}^* \in A)}{M},$$

because $\forall \mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ it is $f_P(\mathbf{X}^*)\mathrm{d}\mathbf{X}^* = f_P(\mathbf{X})\mathrm{d}\mathbf{X}$. It is worth noting here that for calculating the conditional probability distribution it is not necessary to make reference to the so-called *hypothetical repeated sampling principle.* Actually, $\Pr\{\mathbf{X}^* \in A|\mathscr{X}_{/\mathbf{X}}\}$ is *objectively determined* by complete enumeration of $\mathscr{X}_{/\mathbf{X}}$ which once data are observed has a physical existence, and so no hypothetical sampling experiment is referred to in its determination. Since in determining the permutation probability measure in H_0 knowledge of P, or of f_P, is not required, it is to be emphasized that *only the existence of a likelihood is required* by the PT approach (if this existence could not be assumed, no statistical problem would be on the stage). One more relevant consequence of finiteness of $\mathscr{X}_{/\mathbf{X}}$ is that in H_0 permutations $\mathbf{X}^*$ are equally likely conditionally, i.e. $\Pr\{\mathbf{X} = \mathbf{x}|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\mathbf{X}^* = \mathbf{x}|\mathscr{X}_{/\mathbf{X}}\} = 1/M$ if $\mathbf{x} \in \mathscr{X}_{/\mathbf{X}}$ and 0 elsewhere. And so:

- **P.2.** *In H_0 the data set $\mathbf{X}$ is uniformly distributed over $\mathscr{X}_{/\mathbf{X}}$ conditionally.*
- **P.3.** (Uniform similarity of randomized PTs). *Let us assume that the exchangeability condition on data $\mathbf{X}$ is satisfied in H_0, then the conditional rejection probability $\mathbb{E}\{\phi_R(\mathbf{X})|\mathscr{X}_{/\mathbf{X}}\}$ of randomized test $\phi_R = 1$ if $T^o > T_\alpha$, $= \gamma$ if $T^o = T_\alpha$,* and *$= 0$ if $T^o < T_\alpha$, is $\mathbf{X}$-P-invariant for all $\mathbf{X} \in \mathscr{X}^n$ and all $P \in \mathscr{P}$, where: $T^o = T(\mathbf{X})$ is the observed value of T on data $\mathbf{X}$, T_α is the α-sized critical value, and $\gamma = \left[\alpha - \Pr\left\{T^o > T_\alpha|\mathscr{X}_{/\mathbf{X}}\right\}\right] \Big/ \Pr\left\{T^o = T_\alpha|\mathscr{X}_{/\mathbf{X}}\right\}$.*

  For non-randomized PTs such a property is satisfied in the almost sure form for continuous variables and at least asymptotically for discrete variables.

  Determining the critical values T_α of a test statistic T, given the observed data $\mathbf{X}$, in practice presents obvious difficulties. Therefore, it is common to make reference to the associated p-value. This is defined as $\lambda = \lambda_T(\mathbf{X}) = \Pr\{T^* \geq T^o|\mathscr{X}_{/\mathbf{X}}\}$, the determination of which can be obtained by complete enumeration of $\mathscr{X}_{/\mathbf{X}}$ or estimated, to the desired degree of accuracy, by a conditional Monte Carlo algorithm based on a random sampling from $\mathscr{X}_{/\mathbf{X}}$ (Pesarin and Salmaso 2010). For quite simple problems it can be evaluated by efficient computing routines such as those in Mehta and Patel (1983); moreover, according to Mielke and Berry (2007) it can be approximately evaluated by using a suitable approximating distribution, e.g. as within Pearson's system of distributions,

sharing the same few moments of the exact permutation distribution, when these are known in closed form in terms of data $\mathbf{X}$.

The p-value λ is a non-increasing function of T^o and is one-to-one related with the attainable α-value of a test, in the sense that $\lambda_T(\mathbf{X}) > \alpha$ implies $T^o < T_\alpha$, and vice versa. Hence, the non-randomized version can be stated as $\phi = 1$ if $\lambda_T(\mathbf{X}) \le \alpha$, and $= 0$ if $\lambda_T(\mathbf{X}) > \alpha$, for which in H_0 it is $\mathbb{E}\{\phi(\mathbf{X})|\mathscr{X}_{/\mathbf{X}}\} = \Pr\{\lambda_T(\mathbf{X}) \le \alpha|\mathscr{X}_{/\mathbf{X}}\} = \alpha$ for every *attainable* α. Thus, attainable α-values play the role of critical values, and in this sense $\lambda_T(\mathbf{X})$ itself is a test statistic.

- **P.4.** (Uniform null distribution of p-values). *Based on* **P.1**, *if X is a continuous variable and T is a continuous non-degenerate function, then p-value $\lambda_T(\mathbf{X})$ in H_0 is uniformly distributed over its attainable support*.
- **P.5.** (Exactness of permutation tests). *A PT T is exact if its null distribution essentially only depends on exchangeable null deviates* $\mathbf{X}(0)$.
- **P.6.** (Uniform unbiasedness of test statistic T). *PTs for random shift alternatives* $(\Delta \overset{d}{\ge} 0)$ *based on divergence of symmetric statistics of non-degenerate measurable non-decreasing transformations of the data, i.e.* $T^*(\Delta) = S_1[\mathbf{X}_1^*(\Delta)] - S_2[\mathbf{X}_2^*(\Delta)]$, *where* $S_j(\cdot)$, $j = 1, 2$, *are symmetric functions of their entry arguments* $(\cdot)$, *are conditionally unbiased for every attainable* α, *every population distribution* P, *and uniformly for all data sets* $\mathbf{X} \in \mathscr{X}^n$. *In particular*: $\Pr\{\lambda(\mathbf{X}(\Delta)) \le \alpha|\mathscr{X}_{/\mathbf{X}(\Delta)}\} \ge \Pr\{\lambda(\mathbf{X}(0)) \le \alpha|\mathscr{X}_{/\mathbf{X}(0)}\} = \alpha$, *thus p-value in H_1 is stochastically dominated by that in* H_0: $\lambda(\mathbf{X}(\Delta)) \overset{d}{\le} \lambda(\mathbf{X}(0))$.

  An immediate consequence of **P.6** is that, if $\Delta' \overset{d}{>} \Delta$ and so $\lambda(\mathbf{X}(\Delta')) \overset{d}{\le} \lambda(\mathbf{X}(\Delta)) \overset{d}{\le} \lambda(\mathbf{X}(0))$, the permutation p-values of any T are stochastically decreasingly ordered with respect to effect Δ. Without further assumptions, uniform unbiasedness cannot be extended to two-sided alternatives.

  It is worth observing that uniform similarity **P.3** and uniform unbiasedness **P.6** since are at least satisfied for almost all data sets $\mathbf{X}$ under exchangeability in H_0 do not require random sampling from a population. Thus, they also work for selection-bias sampling.
- **P.7.** (The empirical probability measure, EPM). *For each permutation* $\mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$, *the EPM of any* $A \in \mathscr{A}$ *is defined as* $\hat{P}_{\mathbf{X}^*}(A) = \sum_{i \le n} \mathbb{I}(X_i^* \in A)/n$ *which, since* $\forall \mathbf{X}^* \in \mathscr{X}_{/\mathbf{X}}$ *it is* $\sum_{i \le n} \mathbb{I}(X_i^* \in A)/n = \sum_{i \le n} \mathbb{I}(X_i \in A)/n = \hat{P}_{\mathbf{X}}(A)$, *is a permutation invariant function over* $\mathscr{X}_{/\mathbf{X}}$.

  The latter implies that conditioning on $\mathscr{X}_{/\mathbf{X}}$ is equivalent to conditioning on the EPM $\hat{P}_{\mathbf{X}}(A)$, which then is sufficient too.
- **P.8.** (The power of test T). *The (unconditional or population) power of a PT T as a function of Δ, α, T, P, and n is defined as* $W(\Delta, \alpha, T, P, n) = \mathbb{E}_{P^n}[\Pr\{\lambda_T(\mathbf{X}(\Delta)) \le \alpha|\mathscr{X}_{/\mathbf{X}}^n\}]$. *Of course,* $W(\Delta, \alpha, T, P, n) \ge W(0, \alpha, T, P, n) = \alpha$, $\forall \alpha > 0$, *since, in force of* **P.6** *the integrand is* $\ge \alpha$ *for all* $\mathbf{X} \in \mathscr{X}_{/\mathbf{X}}^n$, *all* $P \in \mathscr{P}$, *and all* n.

It is worth noting that **P.8** implies unconditional unbiasedness. It is also to be noted that the power determination of T implies referring to the hypothetical repeated sampling principle.

To introduce the weak consistency property of PTs, stating that "if $\Delta \overset{d}{>} 0$, as $\min[n_1, n_2] \to \infty$ the rejection probability of test T tends to one for all $\alpha > 0$" , let us first consider sequences of related data sets where first n_1 IID values are from $X_1(\Delta) = X + \Delta$ and the other n_2 from $X_2 = X(0) = X$. Such sequences are denoted by $\{\mathbf{X}^{(n)}(\Delta)\}_{n\in\mathbb{N}} = \{[X_{11} + \Delta_1, \ldots, X_{1n_1} + \Delta_{n_1}, X_{21}, \ldots, X_{2n_2}]\}_{(n_1,n_2)\in\mathbb{N}}$. Of course, $\{\mathbf{X}^{(n)}(0) = \mathbf{X}^{(n)}\}_{n\in\mathbb{N}}$ represents sequences in H_0. Besides, we assume that $n \to \infty$ implies $\min[n_1, n_2] \to \infty$.

- **P.9.** (Weak Consistency). *Let X be any population variable and suppose that $\{\mathbf{X}^{(n)}(\Delta)\}_{n\in\mathbb{N}}$ is a sequence of data the first n_1 IID from $(X_1(\Delta), \mathscr{X})$ and independently the other n_2 IID from $(X, \mathscr{X})$. Suppose that the null distribution of X is $P \in \mathscr{P}$, and let $\varphi : \mathscr{X} \to \mathscr{R}^1$ be any non-decreasing and non-degenerate measurable function. Suppose also that*: (a) *the φ-mean $\mathbb{E}_P[\varphi(X)] = \mathbb{E}_P[\varphi(X(0))]$ is finite, i.e. $\mathbb{E}_P[|\varphi(X)|] < +\infty$*; (b) *the φ-mean in H_1 is such that $\mathbb{E}_P[\varphi(X(\Delta))] > \mathbb{E}_P[\varphi(X(0))]$ for every $\Delta \overset{d}{>} 0$*; (c) *the PT is based on $T^* = \frac{1}{n_1}\sum_{i\le n_1}\varphi(X_i^*)$, or on permutationally equivalent statistics. Then, for every $\alpha > 0$*, (a)–(c) *imply that the rejection probability of the PT ϕ, associated with T^*, converges weakly to one as $n \to \infty$.*

It is worth noting that population variable X can be either real, or ordered categorical, and that its transformation $\varphi(X)$ is real, i.e. continuous, discrete, or mixed. As an application of **P.9** we see details for proving consistency of a test based on well-known Cramér–von Mises statistic for one-sided alternatives. Indeed: (1) with $\Delta \overset{d}{>} 0$, $T^*_{CM} = \sum_{i=1}^{n}[\hat{F}_2^*(X_i) - \hat{F}_1^*(X_i)]$, where $\hat{F}_j^*(z) = \sum_i^{n_j}\mathbb{I}(X_{ji}^* \le z)/n_j$, $j = 1, 2$, is permutationally equivalent to $-\sum_{i\le n}\hat{F}_1^*(X_i)/n$, since $\hat{F}_{\mathbf{X}^{(n)}}(t) = [n_2\hat{F}_2^*(t) + n_1\hat{F}_1^*(t)]/n$ is a permutation invariant function; (2) as F_P is bounded, $\mathbb{E}_P(F_P(X))$ is finite; (3) as $\hat{F}_1^*$ is a sample mean, we have that $\Pr\{|\hat{F}_1^*(z) - \hat{F}_{\mathbf{X}^{(n)}}(z)| < \varepsilon|\hat{F}_{\mathbf{X}^{(n)}}\} \longrightarrow 1$, $\forall z \in \mathscr{R}^1$ and $\varepsilon > 0$; (4) $\Delta \overset{d}{>} 0$ implies $\mathbb{E}_P[F_P(X(\Delta))] < \mathbb{E}_P[F_P(X(0))]$. Therefore, since conditions (a)–(c) are satisfied, T^*_{CM} is weakly consistent.

5 Extending Permutation Inference

The non-randomized permutation test ϕ associated with a given test statistic T based on divergence of symmetric functions of the data possesses both conditional unbiasedness and similarity properties, the former **P.6** satisfied by *all population distributions P and all data sets $\mathbf{X} \in \mathscr{X}^n$*, the latter **P.3** satisfied for continuous, non-degenerate variables and *almost all data sets*. These two properties jointly suffice to weakly extend conditional inferences to unconditional or population ones, i.e. for the extension of conclusions related to the specific set of actually observed units (e.g., *drug is effective on the observed units*) to conclusions related to the

population from which units have been drawn (e.g., *drug is effective*). Such an extension is done with weak control of inferential errors. With clear meaning of symbols let us observe:

- (i) for each attainable α and all sample sizes n, the similarity property implies that the power of the test under H_0 satisfies $W(0, \alpha, T, P, n) = \alpha$, because $\Pr\{\lambda(\mathbf{X}(0)) \leq \alpha | \mathscr{X}^n_{/\mathbf{X}}\} = \alpha$ for almost all samples $\mathbf{X} \in \mathscr{X}^n$ and all continuous non-degenerate distributions P, independently of how data are selected;
- (ii) the uniform conditional unbiasedness implies that the unconditional power is $W(\Delta, \alpha, T, P, n) \geq \alpha$ **(P 8)** for all distributions P and, provided that $f_P^{(n)}(\mathbf{X}) > 0$, independently of how data are selected.

As a consequence, if, for instance, the inferential conclusion related to actual data $\mathbf{X}$ is in favor of H_1, so we say that "data $\mathbf{X}$ are evidence of treatment effectiveness on actually observed units," due to (i) and (ii) we are allowed to say that this conclusion is also valid unconditionally for all populations $P \in \mathscr{P}$ such that $f_P^{(n)}(\mathbf{X}) > 0$. Thus, the extended inference becomes "treatment is likely to be effective." The condition $f_P^{(n)}(\mathbf{X}) > 0$ implies that inferential extensions must be carefully interpreted. To illustrate this aspect simply, let us consider an example of an experiment in which only males of a given population of animals are observed. Hence, based on the result actually obtained, the inferential extension from the observed units to the selected sub-population is immediate. Indeed, on the one hand, rejecting the null hypothesis with the actual data means that *data are evidence for a non-null effect of treatment*, irrespective of how data are collected, provided that they are exchangeable in the null hypothesis. On the other hand, if females of that population, due to the selection procedure, have a probability of zero of being observed, then in general we can say nothing reliable regarding them, because it may be impossible to guarantee that the test statistic used for male data satisfies conditional unbiasedness and/or similarity properties for female data as well. For instance, effect may be positive on male and negative on female. In general, *the extension* (i.e., the extrapolation or the inductive generalization) *of any inference to populations which cannot be observed can be formally done only with reference to assumptions that lie outside those that are adopted under the control of experimenters while working on actual data*. For instance, extensions to humans of inferences obtained from experiments on animals essentially require specific hypothetical assumptions.

We observe that for parametric tests, when there are nuisance entities to remove, the extension of inferences from conditional to unconditional can generally be done only if the data are obtained through well-designed sampling procedures applied to the entire target population. When selection-bias data $\mathbf{X}$ are observed and the selection mechanism is not well designed and/or modelled there is no point in staying outside the conditioning on the associated sufficient orbit $\mathscr{X}_{/\mathbf{X}}$ and the related distribution induced by the chosen statistic T. On the one hand this implies adopting the permutation testing principle; on the other, no parametric approach can be invoked to obtain reliable inferential extensions.

References

Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman and Hall, London (1974)

Frosini, V.B.: On some applications of the conditionality principle. Statistica Applicata **3**, 555–568 (1991)

Hoeffding, W.: The large-sample power of tests based on permutations of observations. Ann. Math. Stat. **23**, 169–192 (1952)

Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)

Mielke, P.W., Berry, K.J.: Permutation Methods, A Distance Function Approach, 2nd edn. Springer, New York (2007)

Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data, Theory, Applications and Software. Wiley, Chichester (2010)

Bayesian Statistical Inference: An Overview

Ludovico Piccinato

Abstract

The Bayesian approach for statistical inference is examined, pointing out also the differences among the many Bayesian philosophies. Moreover comments are given about topics where the Bayesian approach seems (at least to Bayesians) more suitable than the alternatives. At last the decision-theoretic approach is shortly discussed.

1 An Historical Outline

In the second half of the nineteenth century the dominant approach to statistical inference was the framework originated by P.S. Laplace, where an honour place was given to Bayes theorem. Both Bayes and Laplace are sometimes mentioned as supporters of a very strict approach to probability and inference: for the classical problem of the *probability of causes*, they would assume an equal probability for the causes, so that, in a modern language, the final probabilities would be proportional to the likelihoods. As shown by authoritative historians, this picture is not correct. Indeed Bayes in his famous 1763 posthumous paper assumed equal prior *predictive* probabilities (that is probabilities of observables) so that the equiprobability of causes was derived as a consequence (Stigler 1982, 1986). The *Principle of Indifference* formulated by P.S. Laplace in 1774 states that the ratio of the final probabilities of two causes A_i and A_j conditional on an event E equals the likelihood ratio $P(E|A_i)/P(E|A_j)$. This amounts to say that the initial probabilities of the causes are equal. But in many places Laplace himself explains that when the cases at hand are not equally possible, one has to subdivide

L. Piccinato (✉)
Sapienza Università di Roma, Rome, Italy
e-mail: ludovico.piccinato@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_5,

or join them to reach a set of equipossible cases. Therefore, even for Laplace, equiprobability is the result of an elaboration, not an aprioristic assumption. It has also been observed (Stigler 1986, p. 135) that in some occasions Laplace explicitly adopted non-uniform priors. Also Karl Pearson in most of his works is clearly sympathetic with the approach based on the so-called *inverse probabilities* (see, e.g., Dale 1991), and his influence was relevant at least until the first decades of the twentieth century. Moreover many authors suggest the use of an uniform prior because it is approximately justified in the case of large sample size, a position that was deepened in modern times.

The break in the tradition is due to the work of Fisher. His main theoretical contribution in this period is Fisher (1922), where Sect. 1 starts with a severe criticism of a famous paper by Pearson (1920). Fisher's offensive, followed after a few years by the well-known contributions of J. Neyman and E.S. Pearson, provoked an eclipse of the Bayesian approach for about three decades (Zabell 1989). This does not mean that in the same period remarkable developments in the Bayesian framework did not occur. On the contrary, in the same period, important works by H. Jeffreys, I.J. Good, F.P. Ramsey and B.de Finetti were published; the point is that the statistical community paid very few attention to such arguments, whose relevance was recognized only many decades later. The change occurred since the 1950s (of the twentieth century) with the work of some scholars, including L.J. Savage, H. Raiffa and R. Schlaifer, D.V. Lindley (Savage 1954, 1962; de Finetti 1959; Ramsey 1926; Lindley 1965). English translations of some key works of de Finetti were provided. In the 1970s several books were published where a definitive setting was given to the Bayesian theory. These include DeGroot (1970), de Finetti (1970), Box and Tiao (1973), Lindley (1972), Berger (1985) and, for the Italian literature, Daboni and Wedlin (1982) and Cifarelli and Muliere (1989). For more bibliographical details see Fienberg (2005); extensive historical information is given in Fienberg (1992) and Fienberg (2006).

Let us assume a standard statistical model, say $\{p(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$, where x is the possible result, θ is the unknown parameter, p is a density or a mass function, and let x_{obs} the observed result. If the goal is to make inferences on the unknown parameter θ, a Bayesian statistician of any century should first of all complete the model adding a probability law (the *prior distribution*) for the parameter, say $\pi(\theta)$. Then he/she can use the celebrated Bayes' formula $\pi(\theta|x_{\text{obs}}) \propto \pi(\theta)p(x_{\text{obs}}|\theta)$ which provides the *posterior* probability distribution for the parameter, i.e. the probability distribution updated with the acquired information. The use of priors is the most evident difference between Bayesian and non-Bayesian methods and a more detailed analysis will be given in the next section.

Automatically, the use of Bayes' formula implies that the values $p(x|\theta)$ with $x \neq x_{\text{obs}}$ have no effect on the analysis. On the contrary the frequentist approach to statistical inference produces conclusions which depend on the whole statistical model, not only on the likelihood function $L(\theta|x_{\text{obs}}) = p(x_{\text{obs}}|\theta)$. In this writer's opinion this aspect, that is the violation of the so-called *Likelihood Principle*, is what mostly moves the frequentist approach away from the Bayesian approach. More comments will be given in Sect. 3.

2 Prior Probabilities

In the first decades of the twentieth century the concept of probability was studied in great depth. One well-known approach sees the probability as a limit of observable frequencies, an interpretation which is not suitable for the Bayesian methodology because of its lack of generality. On the contrary, the probability as a measure of belief in the occurrence of an uncertain event (*subjective* or *personal* probability) has clearly a general applicability. A way to evaluate a probability is a comparison with a standard (see Bertrand 1907, and for a more recent version Lindley 2006). Another approach, developed independently by F.P. Ramsey (1926) and by B. de Finetti (1931) specifies the probability as the fair price of an unitary stake in a bet on an uncertain event. Then de Finetti introduced the principle of *coherence*, i.e. that a subject must avoid bets where he would lose whatever the result and showed that such principle is equivalent to the standard Kolmogorov axioms of probability (not considering the complete additivity, which turns out to be a possible but non necessary choice). Moreover de Finetti formulated the problem of inference as prediction of future results given a partial initial trajectory of the stochastic process of the observations (de Finetti 1937). This formulation does not introduce unknown parameters and replaces the standard notion of random sampling with the concept of *exchangeability*. At a first sight this approach is radically different from the standard one, popularized by Fisher and Neyman and based on the usual statistical models. However the celebrated de Finetti representation theorem shows that exchangeability corresponds to conditional independence so that the procedure based on prior plus likelihood is essentially equivalent to the completely predictive approach, since the assumptions on the process also determine prior and likelihood.

The collaboration between de Finetti and Savage in the 1950s contributed very much to the revival of the Bayesian approach in general, in particular to the acceptance of subjective probabilities. Note that in the Chap. XII of de Finetti (1970), after a short premise about its connection with the predictive approach, the problem of inference is directly treated in the current model-based framework. A common way to respect the original predictive approach by de Finetti is to "justify" the model-based approach through the representation theorem; see, e.g., Dawid (1982). The most systematic treatment in this framework was given by Bernardo and Smith (1994). A claim in favour of the predictivistic approach was presented by Cifarelli and Regazzini (1982) and remarkable methodological researches were conducted under this perspective. For instance, in Regazzini (1999) such approach is explored in a nonparametric context. Less common are treatments oriented to applied problems; the exceptions include Muliere and Petrone (1993) and Spizzichino (2001). A general analysis of de Finetti's work in mathematical statistics cannot be given here, and we refer to Piccinato (1986), Cifarelli and Regazzini (1996), Bernardo (1998) and the references therein. The implementation of the subjectivistic paradigm requires a new interest for the problem of elicitation, i.e. how to put in a probabilistic form the knowledge owned by the experts. Many papers were dedicated to this topic, starting with de Finetti and Savage (1962); a

basic remark is that it is often easier to give a probability to the observables than to unknown parameters. For a recent systematic review see O'Hagan et al. (2006).

A concept by de Finetti which found only a limited acceptance among statisticians was finite additivity (for a deepening see Cifarelli and Regazzini 1996). It is known, however, that complete additivity allows to use properties which hold in the finite problems (for instance, conglomerability) so that its adoption is natural when infinity appears essentially as an approximation of large or unprecised numbers. For special problems, when infinity has its own specific role, resorting to finite additivity can be clarifying also in practical settings (see, e.g., Scozzafava 1984).

In the 1960s the classical argument about the almost irrelevance of the prior in the presence of a significant experimental information (the *Principle of Precise Measurement*) was reconsidered and clarified (Savage 1962; Edwards et al. 1963). This topic has a connection with the recurrent idea of using *noninformative* priors. In the classic period the uniform distribution was often and naively used in this sense, though many authors remarked that the uniformity is not maintained under one-to-one transformations, while on the contrary noninformativity should remain. H.Jeffreys introduced in the 1930s his *invariant rule*, which satisfies this property. The concept of noninformativeness, to take it seriously, has surely very weak bases: a probability distribution always represent an information. This explains why a multiplicity of different proposals were advanced in the years (see Kass and Wasserman 1993). One of these proposals, cautiously named *reference prior*, is based on the idea of minimizing the missing information; it partially extends Jeffrey's rule and is now almost a standard. The proposal originated by a paper by J.M. Bernardo (1979) and was later developed in particular by J. Berger. For more recent treatments see Berger and Bernardo (1992) and Berger et al. (2009). A criticism to the method is that it entails a violation of the Likelihood Principle, since the posterior distribution depends not only on the likelihood function but also on the model (see, e.g., the discussion by Lindley of Bernardo 1979). It could be remarked that this kind of prior (as Jeffreys') is necessarily connected with the model since it is obviously impossible to speak of minimal information in an absolute sense; compatibility with the Likelihood Principle is, however, hold by Bernardo (2005, Sect. 3.6).

The availability of an agreed default rule, where no effort of elicitation is required, suggested an approach which is now called *Objective Bayesian Analysis*. For a comparison of the contrasting arguments see Berger (2006), Goldstein (2006) and the related discussion. Authoritative proponents of the objective approach (Berger et al. 2009) remark that the term "objective" means that the procedure only depends on the model assumed and the data obtained, so that the kind of objectivity is simply the same of the frequentist statistics. There are significant practical and logical differences with a pure subjectivistic approach, but, in the present writer's opinion, these are only variants of a more general Bayesian framework. As mentioned before, I think that the qualification "Bayesian" is due when we assume that any uncertain event has a probability. It is not necessary that there exists a subject who has actually such information. In any case the Bayes theorem explains how to update an information, be it effective or conventional.

In order to simplify the elicitation process many suitable partial formalizations are in use. Among the most known tools there are the conjugate classes of priors. The concept had a systematic treatment by Raiffa and Schlaifer (1961) but the same idea (often limited to the binomial model) appeared many times much before. Until the availability of the MCMC techniques it was often difficult to get the posterior distributions unless the prior was a member of a conjugate class. Using the de Finetti theorem Lindley (1965) represented exchangeable parameters through a hierarchical model. This allowed a very convenient Bayesian treatment of the general linear model (see Lindley and Smith 1972 for a generalization).

In the last decades procedures pointing at conventional priors were suggested and proved useful in applications. For instance, the book (Spiegelhalter et al. 2004) made popular the use of sceptical priors, mainly in a clinical context. Another technique of modelling the prior is the use of power priors, initially proposed in Ibrahim and Chen (2000), that is suitable, for instance, when there are historical data similar to those at hand but not such to justify the assumption of exchangeability (see also De Santis 2007).

Another departure from an ideal subjectivistic practice is the distinction, now very much used, between *design* prior and *analysis* prior. This idea appears from the first time in Tsutakawa (1972) and was developed with various motivations, as the necessity of having a proper prior in the stage of design (while in the stage of analysis an improper prior is often preferred) or of privileging the region of the parameter space which could make the results more interesting (see Etzioni and Kadane 1993; Wang and Gelfand 2002).

3 Statistical Models and Likelihood Principle

In the framework of a standard statistical model the Likelihood Principle has its own strength even without reference to the Bayesian paradigm. Savage, in the discussion of Birnbaum (1962), writes that he came to Bayesian statistics seriously only through recognition of the Likelihood Principle. The issue is, however, controversial; for instance, Cox (2006, p. 47) comments that the principle is convincing in its weak version (two results under the same model are equivalent when the likelihood functions are proportional) but qualifies "less compelling" the strong version (when it is not required that the model is fixed). It is well known that among the merits of Fisher there is the introduction of the likelihood function (Fisher 1922). His attitude about the Likelihood Principle has been largely discussed; for a thorough analysis see Savage (1976). The formal definition is due to Birnbaum (1962), but the argument was already informally in use. Many Bayesian authors stress the relevance of the principle in the context of a Bayesian analysis, see, e.g., Edwards et al. (1963) and Lindley (1972); moreover the likelihood literature is a source of interest for the Bayesian school (we could mention at least Basu 1975, Royall 1997). A definitive treatment is Berger and Wolpert (1988).

A Bayesian attitude is also indirectly favoured by the apparent necessity of using sometimes at least a partial conditioning. One of the most famous examples, the

case of the two laboratories, was published by D.R. Cox (1958). The example shows that a rigid applications of the frequentist rule, which implies an exclusive attention to the long run performances of the statistics, can be untenable, while if one conditions on a suitable ancillary statistic the paradox disappears, as it automatically happens in a Bayesian analysis. This kind of examples took many years to became popular (with the exception of the Bayesian literature), at least in the textbooks. I can just mention that E.L. Lehmann, in the second edition of his classic *Testing Statistical Hypotheses*, added a last chapter where the topic is thoroughly examined and a serious comment on the suitability of the unconditional approach is provided (Lehmann 1986, p. 541): "if repetitions [...] are potential rather than actual interest will focus on the particular event at hand, and conditioning seems more appropriate". Therefore the comparison among the main theories of inference involves more the comparison between choosing a conditioning statistic and choosing a prior distribution, than adopting an objective or a subjective approach. Let us finally mention that recent researches by Bayesian authors about the relationships between the different inferential approaches give a special role just to the conditional frequentist approach (Berger 2003; Bayarri and Berger 2004).

The main advantage of the model-based approach is the possibility of separating the different sources of information, i.e. the pre-experimental information, inbedded in the prior, and the experimental information, inbedded in the likelihood function (in the framework of the model). However, this approach is not completely general for inference problems. Difficulties in finding an agreed specification of the likelihood function were, for instance, considered in Bayarri et al. (1988). In any case, however, a Bayesian can resort to the completely predictive approach in the sense of de Finetti, though this could force to reformulate inferential problems.

4 The Development of Bayesian Methodology

Many hints to the development of Bayesian methodology were provided by the existing frequentist methodology: problems having a solution in a non-Bayesian approach had to be revised and reformulated. I shall comment some examples.

One of these themes is *robustness*, that was initially considered in the Bayesian literature mainly in relation to the choice of the prior. Instead of considering a single prior, classes of priors were taken into account in order to check the resulting differences. Beyond parametric classes, attention was drawn also to nonparametric or to partially nonparametric classes, as the class of monotone distributions, of symmetric distributions, or contaminated distributions, quantile classes and so on (for reviews see Berger et al. 1996, Ríos Insua and Ruggeri 2000). This rich literature allowed to move from mathematical convenience to much more realistic formulations of prior uncertainty. The proposal of interactive procedures (as in Liseo et al. 1996) was a further step in this direction.

Another topic inherited by the frequentist statistic is the issue of *model testing and selection*. In a controversial paper Box (1980a,b) claimed that the Bayesian analysis is fully adequate within a given model but is not useful for model criticism.

His proposal for model criticism is based on the prior predictive distribution and has a clear frequentist flavour, together with an analogy with the classical p-value. This proposal suggested many developments in different directions. From one hand, letting aside the traditional criticism to the theory of significance from a Bayesian viewpoint (a seminal paper is Berger 1986), new concepts of Bayesian p-values were introduced, with application to the case of composite hypotheses and to the model criticism (Bayarri and Berger 2000). When the goal is to choose one model many authors suggest an explicit decision setting; see, for instance, San Martini and Spezzaferri (1984), Key et al. (2001), Walker et al. (2001), Barbieri and Berger (2004). The most natural Bayesian approach to compare many models, when one is considered "true" (the so called M-closed setup), is however to attach probabilities to every model, in order to account for model uncertainty, and proceed with the standard probability rules. A general exposition of the Bayesian model averaging is Hoeting et al. (1999). An alternative path is the use of Bayes factors for comparing models without assigning prior probabilities to the model themselves. This was, for instance, proposed by O'Hagan in the discussion of Box (1980b). Problems associated with the Bayes factors should, however, be considered; see Lavine and Schervish (1999) about their use as measures of evidence and Carota and Parmigiani (1996) about their use with nonparametric models. For general treatments refer to Kass and Raftery (1995) and Berger (1999). In the comparison of models it may be desirable to assign improper priors to the parameters of each model. Apart from special situations (e.g., Consonni and Veronese 1991), a general solution is resorting to the so-called partial Bayes factors; for different proposals and discussions see Berger and Pericchi (1996), O'Hagan (1995) and De Santis and Spezzaferri (1997). For the general topic of model selection, including also the assumption of a Model-open setting, that is when it is not assumed that the set of models contains the "true" model, see Racugno (1997), Lahiri (2001), Kadane and Lazar (2004) and Clyde and George (2004).

As another example, let us mention *nonparametric inference*. The mathematical modeling of the problem requires the use of probability measures on function spaces so that the practical understanding of the prior assumptions is quite demanding and a Bayesian treatment was delayed for a long time. Lindley (1972, p. 66) wrote "this is a subject about which the Bayesian method is embarrassingly silent". This was true, at those times, although in a very short paper, many years before, de Finetti (1935) outlined the issue in a Bayesian framework (comments on this in Cifarelli and Regazzini 1996). A turning point was the approach by Ferguson (1973) through the so-called Dirichlet process, which gave rise to many of the contemporary researches. Many different extensions and alternatives were since then proposed (see, e.g., Walker and Muliere 1997, Lijoi and Prünster 2000).

At last we mention some problems that the Bayesian approach can handle in a particularly easy way, differently from the other approaches. These include the elimination of *nuisance parameters*, the possibility of a direct treatment of *prediction problems*, the possibility of a complete treatment of the *design of experiments*. Let us suppose that the parameter θ is a vector, say $\theta = (\lambda, \gamma)$, and that the inference concerns only the component λ. The likelihood function depends

of course on both the components, but, given the posterior distribution, we can get a posterior distribution for the parameter of interest alone by a simple marginalization, that is $\pi(\lambda|x_{\text{obs}}) = \int \pi(\lambda, \gamma|x_{\text{obs}})\text{d}\gamma$. For a recent review see Liseo (2006).

A problem of prediction is characterized by a statistical model $(q(y|\theta), y \in \mathscr{Y}, \theta \in \Theta)$, for the future result Y, where the "true" parameter θ is the same of the statistical model of the observation X. Under the only assumption of the independence of X and Y (for a given θ) it is impossible to represent how the knowledge of X provides information on Y. On the contrary, the introduction of a prior distribution $\pi(\theta)$ for the parameter allows us to calculate the conditional distribution of Y given X, that is $m(y|x_{\text{obs}}) = \int q(y|\theta)\pi(\theta|x_{\text{obs}})\text{d}\theta$ which is the most natural base for a prediction of the value y. A general reference across the different approaches, is Geisser (1993).

In a problem of design of an experiment we have a class $\mathscr{E}$ of possible experiments, which can differ, for instance, for size of the sample, sequential stopping rule, choice of the controlled variables and so on. Any choice $e \in \mathscr{E}$ will get an evaluation depending in general on the result x and the parameter θ, both not known in advance. Under these conditions, general methods of eliminating θ without an integration with respect to a prior distribution are unreasonable or unavailable, unless there are particular patterns as it may occur with linear models (Kiefer's theory). The case of linear models were reformulated also in a Bayesian setting, see, e.g., Smith and Verdinelli (1980) and Giovagnoli and Verdinelli (1983). A particular problem, of a great practical importance, is the determination of an optimal sample size. The diffusion of the Bayesian approach produced a lot of new methods; a starting point for the more recent researches in this field is the issue 2, 1997 of the journal *The Statistician*, entirely devoted to the subject; see also De Santis (2006) for a robust approach. Note that the choice of a design s primarily a decision problem, though the final goal could be an inferential statement. An excellent framework also for this particular problem is therefore given in the classic text by Raiffa and Schlaifer (1961). For further reviews and treatments always in a Bayesian setting see Chaloner and Verdinelli (1995) and Piccinato (2009).

5 Relations with the Decision-Theoretic Approach

The decision-theoretic approach has many merits in clarifying the differences between the approaches. One may ask whether the reformulation in decision-theoretic terms does not modify or restrict the aims of inference. This is surely not true for the Neyman–Pearson–Wald school, because in that case the idea of optimization is intrinsic to the theory. It is well known that many times Neyman explained how inductive reasoning is often impossible since it involves events lacking a probability; inductive behaviour, i.e. optimizing the long run performance of procedures, would be instead the operational solution (see, e.g., Neyman 1957). For the Bayesian approach the situation is different, since both paths are possible; either completing the model involving also a specification of the available terminal

acts with the corresponding utilities/losses or performing a purely probabilistic analysis, without a formal implication of any decision.

If we explicitly adopt a complete decision setting, it is natural that the Bayesian approach aims at minimizing the expected loss of the terminal actions conditional on a specific result, while the frequentist approach aims at minimizing the risk of a procedure unconditionally on the result but conditionally on the unknown parameter. Wald (1950) proved that there is a strong connection between the two optimalities (the complete class theorem). Loosely speaking, the theorem shows that any reasonable decision is formally Bayesian and vice versa. In the Wald's approach prior probabilities are only weighting devices, but this result can be commented as a partial conciliation between the two approaches (see, e.g., de Finetti 1951, Raiffa and Schlaifer 1961, p. 16). The calculation of the risk requires, however, an integration on the sample space, and this is a violation of the likelihood principle; this can imply contradictions with the basic goal of minimizing losses for terminal acts (see, e.g., Piccinato 1980). A good long run performance is in itself a sensible characteristic for a statistical procedure but it should not be achieved by omitting to take into account the actual result, when available.

6 Final Remarks

The aim of the present paper is to outline the many theories and proposals framed in the Bayesian setting (for further information see Berger 2000). We hold that all this is a richness of the approach and does not preclude its fundamental unitariety, based on the formal representation of the process of learning from experience. Limitations of space and knowledge prevent here any hope of completeness. It is worth noting that the development of simulation methods allows now to deal with complex models, with thousands of parameters, as it occurs in the modern applications in genomics and in environmental analysis (see, e.g., Chen et al. 2010), while in the past mathematical tractability played a serious limiting role. In the next future, overviews of the Bayesian approach will surely focus much more on these aspects.

References

Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. Ann. Stat. **32**, 870–897 (2004)

Basu, D.: Statistical information and likelihood. Sankhyā A **37**, 1–71 (1975)

Bayarri, M.J., Berger, J.O.: P-values for composite null models (with discussion). J. Am. Stat. Assoc. **95**, 1127–1170 (2000)

Bayarri, M.J., Berger, J.O.: The interplay of Bayesian and frequentist analysis. Stat. Sci. **19**, 58–80 (2004)

Bayarri, M.J., DeGroot, M.H., Kadane, J.B.: What is a likelihood function? In: Gupta, S.S., Berger, J.O. (eds.) Statistical Decision Theory and Related Topics, vol. 1, pp. 3–16. Springer, New York (1988)

Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (1985)

Berger, J.O.: Are P-values reasonable measures of accuracy? In: Francis, I.S. et al. (eds.) Pacific Statistical Congress. Elsevier, Amsterdam (1986)
Berger, J.O.: Bayes factors. In: Kotz, S., Read, C,B., Banks, D.L. (eds.) Encyclopedia of Statistical Sciences. Update, vol. 3, pp. 20–29. Wiley, New York (1999)
Berger, J.O.: Bayesian analysis: a look at today and thoughts of tomorrow. In: Raftery, A.E., et al. (eds.) Statistics in the 21th Century, pp. 275–290. Chapman and Hall/CRC, Boca Raton (2000)
Berger, J.O.: Could Fisher, Jeffreys and Neyman have agreed on testing? Stat. Sci. **18**, 1–32 (2003)
Berger, J.O.: The case for objective Bayesian analysis. Bayesian Anal. **3**, 385–402 (2006)
Berger, J.O., Bernardo, J.M.: On the development of reference priors. In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 4. Clarendon Press, Oxford (1992)
Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. **91**, 109–122 (1996)
Berger, J.O., Wolpert, R.: The Likelihood Principle, 2nd edn. Institute of Mathematical Statistics, Hayward (1988)
Berger, J.O., Betrò, B., Moreno, E., Pericchi, L.R., Ruggeri, F., Salinetti, G., Wasserman, L. (eds.): Bayesian Robustness. Institute of Mathematical Statistics, Hayward (1996)
Berger, J.O., Bernardo, J.M., Sun, D.: The formal definition of reference priors. Ann. Stat. **37**, 905–938 (2009)
Bernardo, J.M.: Reference posterior distributions for Bayesian inference (with discusion). J. Roy. Stat. Soc. Ser. B **41**, 113–147 (1979)
Bernardo, J.M.: Bruno de Finetti en la Estadística Contempornea. In: Rìos, S. (ed.) Historia de la Matmatica en el siglo XX, pp. 63–80. Real Academia de Ciencias, Madrid (x 1998)
Bernardo, J.M.: Reference analysis. In: Dey, D.K., Rao, C.R. (eds.) Handbook of Statistics, pp. 17–90. Elsevier, Amsterdam (2005)
Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, Chichester (1994)
Bertrand, J.: Calcul des Probabilités. Chelsea, New York (1907)
Birnbaum, A.: On the foundations of statistical inference. J. Am. Stat. Assoc. **57**, 269–306 (1962)
Box, G.E.P.: Sampling and Bayes' inference in scientific modelling and robustness (with discussion). J. Roy. Stat. Soc. Ser. A **143**, 383–430 (1980a)
Box, G.E.P.: Sampling and Bayes' inference and robustness in the advancement of learning (extended abstract with discussion). In: Bernardo, J.M., et al. (eds.) Bayesian Statistics. University Press, Valencia (1980b)
Box, G.E.P., Tiao, G.C.: Bayesian inference in statistical analysis. Addison-Wesley, Reading (1973)
Carota, C., Parmigiani, G.: On Bayes factors for nonparametric alternatives. In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 5, pp. 507–511. Clarendon Press, Oxford (1996)
Chaloner, K., Verdinelli, I.: Bayesian experimental design: a review. Stat. Sci. **10**, 273–304 (1995)
Chen, M-H, Dey, D.K., Mller, P., Sun, D., Ye, K. (eds.): Frontiers of Statistical Decision Making and Bayesian Analysis. In Honor of James O. Berger. Springer, New York (2010)
Cifarelli, M., Muliere, P.: Statistica Bayesiana. G. Iuculano, Pavia (1989)
Cifarelli, D.M., Regazzini, E.: Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability. In: Koch, G., Spizzichino, F., (eds.) Exchangeability in Probability and Statistics. North Holland, Amsterdam (1982)
Cifarelli, D.M., Regazzini, E.: de Finetti's contribution to probability and statistics. Stat. Sci. **11**, 253–282 (1996)
Clyde, M., George, E.J.: Model uncertainty. Stat. Sci. **19**, 81–94 (2004)
Consonni, G., Veronese, P.: Bayes factors for linear models with improper priors. In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 4, pp. 587–594. Clarendon Press, Oxford (1991)
Cox, D.R.: Some problems connected with statistical inference. Ann. Math. Stat. **29**, 357–372 (1958)
Cox, D.R.: Principles of Statistical Inference. Cambridge University Press, Cambridge (2006)
Daboni, L., Wedlin, A.: Statistica. Un'introduzione alla statistica neo-bayesiana. UTET, Torino (1982)

Dale, A.I.: A History of Inverse Probability. From Thomas Bayes to Karl Pearson. Springer, New York (1991)
Dawid, A.P.: Intersubjective statistical models. In: Koch, G., Spizzichino, F. (eds.) Exchangeability in Probability and Statistics. North-Holland, Amsterdam (1982)
de Finetti, B.: Sul significato soggettivo della probabilità. Fundamenta Mathematicae 17 (1931). Reprinted in: Opere Scelte, vol. I, pp. 191–222. Cremonese, Roma (2006)
de Finetti, B.: Il problema della perequazione. Atti della Soc Ital per il Progresso delle Scienze, XXIII riunione (1935). Reprinted in: Opere Scelte, vol. I, pp. 287–288, Cremonese, Roma (2006)
de Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Ann. Inst. H. Poincar **7**, 1–68 (1937) [English translation in Studies in Subjective Probability (Kyburg, H.E. and Smokler, H.E. eds.), Wiley, New York, 1964]
de Finetti, B.: L'opera di Abraham Wald e l'assestamento concettuale della statistica matematica moderna. Statistica **11**, 185–192 (1951)
de Finetti, B.: La probabilità e la statistica nei rapporti con l'induzione secondo i diversi punti di vista. Cremonese, Roma (1959) (English translation in de Finetti, B.: Probability, Induction and Statistics. Wiley, London, 1972)
de Finetti, B.: Teoria della probabilità. Einaudi, Torino (1970) (English translation in Theory of Probability, Wiley, New York, 1974)
de Finetti, B., Savage, L.J.: Sul modo di scegliere le probabilità iniziali. In: Sui Fondamenti della Statistica. Biblioteca di Metron, pp. 81–154. Università di Roma, Roma (1962)
DeGroot, M.H.: Optimal Statistical Decisions. McGraw-Hill, New York (1970)
De Santis, F.: Sample size determination for robust Bayesian analysis. J. Am. Stat. Assoc. **101**, 278–291 (2006)
De Santis, F.: Using historical data for Bayesian sample size determination. J. Roy. Stat. Soc. Ser. A **70**, 95–113 (2007)
De Santis, F., Spezzaferri, F.: Alternative Bayes factors for model selection. Can. J. Stat. **25**, 503–515 (1997)
Edwards, W., Lindman, H., Savage, L.J.: Bayesian statistical inference for psychological research. Psychol. Rev. **70**, 193–242 (1963)
Etzioni, R., Kadane, J.B.: Optimal experimental design for another's analysis. J. Am. Stat. Assoc. **88**, 1404–1411 (1993)
Ferguson, T.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
Fienberg, S.E.: A brief history of statistics in three and one-half chapters. Stat. Sci. **7**, 208–225 (1992)
Fienberg, S.E.: A "Bayesian classics" reading list. ISBA Bull. **12**, 9–14 (2005)
Fienberg, S.E.: When did Bayesian inference become "Bayesian"? Bayesian Anal. **1**, 1–40 (2006)
Fisher, R.A.: On the mathematical foundations of theoretical statistics. Phil. Trans. R. Soc. Ser. A 309–368 (1922). Reprinted in: Fisher, R.A.: Contributions to Mathematical Statistics. Wiley, New York (1950)
Geisser, S.: Predictive Inference: Asn Introduction. Chapman and Hall, London (1993)
Giovagnoli, A., Verdinelli, I.: Bayes D-optimal and E-optimal block designs. Biometrika **79**, 695–706 (1983)
Goldstein, M.: Subjective Bayesian analysis: principles and practice. Bayesian Anal. **1**, 403–420 (2006)
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.: Bayesian model averaging: a tutorial (with discussion). Stat. Sci. **14**, 382–417 (1999)
Ibrahim, J.G., Chen, M.H.: Power prior distributions in regression models. Stat. Sci. **15**, 46–60 (2000)
Kadane, J.B., Lazar, N.A.: Methods and criteria for model selection. J. Am. Stat. Assoc. **99**, 279–290 (2004)
Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)
Kass, R.E., Wasserman, L.: The selection of prior distributions by formal rules. J. Am. Stat. Assoc. **91**, 1343–1370 (1993)

Key, J.T., Pericchi, L.R., Smith, A.M.F.: Bayesian model choice: What and why? In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 6. Clarendon Press, Oxford (2001)
Lahiri, P. (ed.): Model Selection. Institute of Mathematical Statistics, Beachwood (2001)
Lavine, M., Schervish, M.J.: Bayes factors: what they are and what they are not. Am. Statististician **53**, 119–122 (1999)
Lehmann, E.L.: Testing Statistical Hypotheses, 2nd edn. Wiley, New York (1986)
Lijoi, A., Prünster, I.: Models beyond the Dirichlet process. In: Hjort, N.L., et al. (eds.) Bayesian Nonparametrics, pp. 80–136, Cambridge University Press, Cambridge (2000)
Lindley, D.V.: Probability and Statistics. Cambridge University Press, Cambridge (1965)
Lindley, D.V.: The estimation of many parameters. In: Godambe, V.P., Sprott, D.A., (eds.) Foundations of Statistical Inference pp. 435–447, Holt Rinehart and Winston, Toronto (1971)
Lindley, D.V.: Bayesian Statistics, A Review. SIAM, Philadelphia (1972)
Lindley, D.V.: Understanding Uncertainty. Wiley, Hoboken (2006)
Lindley, D.V., Smith, A.M.F.: Bayes estimates for the linear model. J. Roy. Stat. Soc. Ser. B **34**, 1–18 (1972)
Liseo, B.: The elimination of nuisance parameters. In: Dey, D., Rao, C.R. (eds.) Handbook of Statistics, vol. 25, pp. 193–219. Elsevier, Amsterdam (2006)
Liseo, B., Petrella, L., Salinetti, G.: Robust Bayesian analysis; an interactive approach. In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 5, pp. 661–666. Clarendon Press, Oxford (1996)
Muliere, P., Petrone, S.: A Bayesian predictive approach to sequential search for an optimal dose, parametric and nonparametric models. J. Ital. Stat. Soc. **3**, 349–364 (1993)
Neyman, J.: "Inductive behavior" as a basic concept of philosophy of science. Rev. Int. Stat. **25**, 7–22 (1957)
O'Hagan, A.: Fractional Bayes factors for model comparisons. J. Roy. Stat. Soc. Ser. B **57**, 99–138 (1995)
O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakov, T.: Uncertain Judgements. Eliciting Experts' Probabilities. Wiley, Chichester (2006)
Pearson, K.: The fundamental problem of practical statistics. Biometrika **13**, 1–16 (1920)
Piccinato, L.: On the orderings of decision functions. Symposia Mathematica (Istituto Nazionale di Alta Matematica) **XXV**, 61–70 (1980)
Piccinato, L.: de Finetti's logic of uncertainty and its impact on statistical thinking and practice. In: Goel, P.K., Zellner, A. (eds.) Bayesian Inference and Decision Techniques. Essays in Honor of Bruno de Finetti pp. 13–30, North Holland, Amsterdam (1986)
Piccinato, L.: Metodi per le Decisioni Statistiche, 2nd edn. Springer-Italia, Milano (2009)
Racugno, W. (ed.): Proceedings of the Workshop on Model Selection. Pitagora, Bologna (1997)
Raiffa, H., Schlaifer, R.: Applied Statistical Decision Theory. MIT Press, Cambridge (1961)
Ramsey, F.P.: Truth and probability. In: The Foundations of Mathematics and Other Logical Essays. Kegan, London (1926). Reprinted in: Studies in Subjective Probability (Kyburg, H.E., Smokler, H.E. eds.) Wiley, New York (1964)
Regazzini, E.: Old and recent results on the relationship between predictive inference and statistical modelling either in nonparametric or parametric form (with discusssion). In: Bernardo, J.M., et al. (eds.) Bayesian Statistics 6, pp. 571–588. Clarendon Press, Oxford (1999)
Ríos Insua, D., Ruggeri, F. (eds.): Robust Bayesian Analysis. Springer, New York (2000)
Royall, R.: Statistical Evidence. A Likelihood Paradigm. Chapman and Hall, London (1997)
San Martini A., Spezzaferri, F.: A predictive model selection criterion. J. Roy. Stat. Soc. Ser. B **46**, 296–303 (1984)
Savage, L.J.: The Foundations of Statistics. Wiley, New York (1954) (2nd edn., Dover, New York, 1972)
Savage, L.J.: Subjective probability and statistical practice. In: Barnard, G.A., Cox, D.R. (eds.) The Foundations of Statistical Inference. Methuen, London (1962)
Savage, L.J.: On rereading R.A. Fisher. Ann. Stat. **4**, 441–500 (1976)
Scozzafava, R.: A survey of some common misunderstandings concerning the role and meaning of finitely additive probabilities in statistical inference. Statistica **44**, 21–45 (1984)

Smith, A.M.F., Verdinelli, I.: A note on Bayes designs for inference using a hierarchical linear model. Biometrika **7**, 613–619 (1980)
Spiegelhalter, D.G., Abrams, K.R., Myles, J.P.: Bayesian approaches to clinical trials and health-care evaluations. Wiley, New York (2004)
Spizzichino, F.: Subjective Probability Models for Lifetimes. Chapman and Hall/CRC, Boca Raton (2001)
Stigler, S.M.: Thomas Bayes' Bayesian inference. J. Roy. Stat. Soc. Ser. A **145**, 250–258 (1982)
Stigler, S.M.: The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge (1986)
Tsutakawa, R.K.: Design of Experiment for Bioassay. J. Am. Stat. Assoc. **67**, 584–590 (1972)
Wald, A.: Statistical Decision Functions. Wiley, New York (1950)
Walker, S.G., Muliere, P.: Beta-Stacy processes and a generalisation of the Pólya-urn scheme. Ann. Stat. **25**, 1762–1780 (1997)
Walker, S.G., Gutierrez-Peña, E., Muliere, P.: A decision theoretic approach to model averaging. Statistician **50**, 31–39 (2001)
Wang, F., Gelfand, A.E.: A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Stat. Sci. **17**, 193–208 (2002)
Zabell, S.: R.A.Fisher on the history of inverse probability (with discussion). Stat. Sci. **4**, 247–263 (1989)

Part II

Historical Issues in Economic and Socio-Demographic Studies

The Long Journey of Italian Statistics on International Migration

Corrado Bonifazi

Abstract

The last 150 years of Italian history include all the main steps of the evolution of a national migration system. In fact, for almost a century Italy had been one of the most important countries of emigration in the world, while nowadays it has become one of the favorite destinations of international migration flows. The paper is an attempt to describe the main changes in Italian statistics on international migration, highlighting the relations between these changes and changes in migration policies and migration trends.

1 Introduction

The history of the first 150 years since unification saw Italy go through all the main steps that a national migration system can meet during its evolution. Just think that, for almost 100 years, Italy has been amongst the main countries of emigration in the world, though now it has become one of the main destinations for international migration. From the beginning of the new State, the importance of migration has led to an attempt by official statistics to measure its intensity. Since then, Italian statistics on international migration have experienced profound changes in sources, definitions, and methods of data collection. These changes are closely intertwined, on the one hand, to the evolution and dynamics of migration flows and, on the other, to political choices in the migration field. In fact, it is well known that in every country statistical information on the phenomenon tends to focus on specific categories of migrants subject to political interest and, therefore, to move its attention according to the changes in needs and demands of national

C. Bonifazi (✉)
Institute for Research on Population and Social Policies (IRPPS-CNR), Roma, Italy
e-mail: c.bonifazi@irpps.cnr.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_6,

policy (Kritz 1987). We will try to describe this process, aiming to grasp some of its most important aspects and taking into consideration the four main periods in which we can divide the history of Italian migration: the period of mass emigration from 1861 to the First World War; the inter-war period; the period of European labor migration (1946–1975); the period of mass immigration.

2 Italian Migration Statistics During the Mass Emigration Period

The birth and development of Italian migratory statistics should be included within an international context in which, already before 1861, the problem of collecting comparable information about such a social dynamic, which was growing in importance on the world stage, had already been raised. In fact, as early as 1853, the first International Statistical Congress in Brussels, in its conclusions, identified some guidelines for collecting comparable data for the measurement of international migration, particularly the trans-oceanic types (CGS 1853). The central point of the proposal envisaged the creation of municipal registers of the emigrants in order to gather information on those who moved to another country and alongside such registers, in order to measure the incoming traffic, registers of immigrants. Gathering information at ports of departure and arrival on emigrants and immigrants was recommended as a control strategy. Therefore, since then, two key instruments that are still essential in the measurement of migration were identified: the population registers and statistics on departures and arrivals.

In the early post-unification years, migration certainly was not a priority in a country that was yet to be built and neither was for the newly formed Office of General Statistics that belonged to the Ministry for Agriculture, Industry and Commerce (MAIC). In the first census of 1861 there was, however, information collected on international mobility. In particular, the birthplace was considered to distinguish people born within the kingdom from those born in foreign countries. In addition, information was collected on seasonal migrations, distinguishing between those that occurred within unified Italy and those that moved abroad.

The prevailing models of mobility were still the traditional ones related to agricultural activities in the Po Valley and the Roman countryside or to transhumance. The extent of emigration to other countries was small. Even if some migratory flows had already begun before unification and in some local areas international migration had already become an important element in the economic family strategy. Overall, however, Italy was still far from reaching those continental and extra-continental migration flows that will characterize the latter part of the century. Just consider that in neighboring France, the census of 1861 showed 76,000 Italians, next to 85,000 Germans and 204,000 Belgians, or that the US Census of 1860 counted less than 12,000 Italians when the Irish were already 1.6 million, the Germans nearly 1.3, and the British almost 600,000.

The first decade of unification represented an important moment in the evolution and transformation of Italian migration, with the gradual emergence of new patterns of mobility, the enlargement of emigration area, and the growth of migration

flows. Emigration thus began to take on an increasingly important role in the life of the country. On a political level, such an increased importance resulted in a debate that was most vigorous. On the one side there were those who feared the consequences of this phenomenon and wanted it to be controlled to avoid loss of human resources for the country and on the other, whose who considered it an inevitable result of the economic and social transformations that were taking place (Marucco 2001). The comparison also moved quickly onto a question of numbers. The first census of Italians abroad was completed in 1871. The survey proved itself largely inadequate to provide an accurate understanding of the phenomenon and led to estimate between 432,000 and 478,000 units "the approximate number of Italian residents or visiting in other countries on the night of December 31, 1871" (SGRI 1874, II). In 1871 the results of a semi-official survey by Leone Carpi were also published. Carpi had some collaboration from the Ministry of the Interior and the Foreign Office. In 1872 the Statistical Council urged for the establishment of an emigration statistics (MAIC 1872) which, after various studies, was launched in 1876. The measurement of the phenomenon was seen as an essential tool to assess the causes and effects and to "put an end to the 'controversy over the figures' around which, until then, proponents and opponents of the Exodus had clashed" (Marucco 1996, p. 155). The debate within the Statistical Council concerned, first of all, the very basic position on migration. Proponents of migration were well represented by Luigi Bodio, Secretary of the Board since 1872 and subsequently Director of the Directorate of Statistics, who considered migration a fact of life and a right that must be recognized to all. Opponents were represented by Giovanni Florenzano, author of an emigration statistics for the Province of Naples published in 1874. He, in contrast, saw the migration outflow as a serious detriment to the agricultural economy of the country.

This diversity of positions was also on a more strictly technical level. The line taken by Bodio, which then prevailed, was to point to official statistics, "conducted according to scientific methods and carried out with proper tools, quite willing to limit its aspirations rather than to venture into uncertain terrain" (Marucco 2001, p. 64). Florenzano feared that such a situation would hide the real intensity of the phenomenon, which would result in preventing or delaying government intervention to discourage emigration. Furthermore the first results of the new survey counted about 108,000 units as the outflow in 1876, while the estimates by Carpi counted nearly 152,000 units in 1873 (Bodio 1877). Apart from the political factors, we find in such a diversity of positions the main conceptual node that still characterizes the debate on migration statistics, with the confrontation between the needs and limitations of official statistics on the one hand, and knowledge needs, on the other. In this sense the position of Bodio, careful in identifying as objectively as possible measurement criteria, contains the fundamental character of a modern statistical survey which must be based on reliable data and whose limits of coverage and reliability are particularly clear.

However, the definition of the criteria of a survey, that through successive changes and adjustments accompanied the entire history of Italian emigration, was anything but linear. On 2 March 1874, the Board of Statistics and the Advisory

Commission on pension funds and work, in a joint meeting, identified a number of requirements for the statistics on emigration (MAIC-DS 1880). It is a long list of needs, many of which proved to be impossible to achieve. In fact, the initial path of the source is characterized by a gradual adjustment of its ambitious knowledge goals to the feasible statistical tools. Ultimately the source, up until the changes made in the early twentieth century, considered migrants as emigrating people in poor economic conditions and based the collection of data on clearances granted by the municipalities for the issuing of passports, supplemented by other information. The most apparent limit of the survey, very clear from the beginning to the Directorate of Statistics, was the obvious discrepancy between the number of clearances and passports granted and actually used. This was also at a time when the possession of such a document was not essential to expatriation.

However, beyond this and other limits of the survey, it must be considered that the characteristics of Italian emigration from the start of the phenomenon and the deficiencies of the administrative system made it difficult to use other methods of data collection. In the Italian case, in fact, emigration was initially mainly directed to other European countries, thus making it less meaningful to a system based on boarding lists. This was unlike the situation in other countries where the bulk of the phenomenon was for departures toward North and South America or took place within national borders. The choice of creating registers of emigrants, as suggested in the conclusions of the International Statistical Congress of 1853, would then have been impractical given the difficulties to set up an effective system of population registers, whose creation had been expected as early as 1862 (Marucco 1996) but whose practical realization did not take place until 1929.

The last quarter of the nineteenth century witnessed a considerable growth in the migratory exodus, within an increasingly favorable framework of the phenomenon: for the first time the expatriates in 1887 exceeded 200,000 units and in 1900 reached almost 353,000. On the side of the push factors, the start of the demographic transition, the development of new economic activities, and the crisis of large sections of the traditional economy led to an increase in the need to emigrate. Furthermore emigration soon came to be the most effective and direct way of improving the quality of life for ever larger segments of the population, while the effective action by a range of actors was stronger (especially the shipping companies and migration agents) and created a real economy of migration. Then, the attractive factors were no less intense. In European countries, migrant workers carried out integrative and substitutive jobs in those areas that were deserted by the local workforce, especially in agriculture and in large public works (Bade 2001). On the other side of the Atlantic, countries promoted immigration to promote their development, encouraged by the extraordinary improvement in shipping and the lowering in the cost of crossing (Bade 2001).

The migration issue entered the Italian political debate with even greater force. Though, for the first law on the matter, we must wait until 1888, after a lively parliamentary debate and characterized by an evident conflict of interests between the agrarian South and the shipping Companies (Ostuni 2001). The law that would definitively regulate the great outflow would see the light of day only

in 1901, and was essentially a measure of compromise between the needs for protection and promotion of emigration. The law would foresee the creation of the General Commissariat for Migration (CGE), whose director was Luigi Bodio, and contained the first official definition of the emigrant, at least as far as ocean flows were concerned. With the arrival of the twentieth century, out-migration grew tumultuously: in 1901, 533,000 expatriated and in subsequent years such a phenomena reached levels that would never be reached again, with a maximum of almost 873,000 expatriates in 1913. The focus had now passed to the promotion of emigration, also thanks to the CGE. The macroeconomic effects of migration, through remittances, had meanwhile reached a remarkable size, ensuring that the major economic changes of the Giolitti period come to be in a situation of significant trade deficit coverage (Sori 1979).

Even from the statistical point of view the new century brought important changes. From 1904, the collection of data on expatriates was based on the records of passports held by the District Offices of the Ministry of Interior, which represented a definite improvement (CGE 1926). There remained, however, wide margins of difference between those statistics and the intensity of the phenomenon (CGE 1926). The issue of a passport, in fact, did not necessarily imply migration. Transfers without a passport grew due to the number of states that did not require such a document. Hence, it was not always possible to determine with accuracy the country of destination and, given the 3 years of duration of a passports' validity, more movements abroad were possible for the same person. The CGE too began its own independent survey of ocean flows in 1902, based on the boarding lists and the data was collected on both outward and return movements. The latter formed the basis for statistics on returnees, which would fill a significant gap in a country where return migration and circular mobility have always held a great importance in ocean flows. The major limits of the CGE survey are related to the exclusion of those who traveled in a different class from third class, for those who departed from foreign ports, and those who worked on ships during the journey (CGE 1926).

The arrival of the First World War marked the end of the first globalization and of a time when international migration had been an essential element in the functioning and development of the world economy. After the war, as we shall see in the next section, all the coordinates of the issue radically changed and a new era for Italian and international migration was born.

3 The Inter-War Period

With the conclusion of the conflict, the movements of people were subjected to increasingly stringent controls and restrictions by the states of destination, while the global economy had difficulty in recovering pre-war growth rates and, in 1929, would enter one of its deepest crises. The transoceanic flows were reduced greatly and, in particular, Italian emigration suffered, being subjected to particularly stringent constraints in the United States which, before the conflict, had become the main destination. Also migration flows between European countries met with

a sharp decline as a result of poor post-war economic conditions and political restrictions. Only France continued to have a policy of attracting flows, at least until the effects of the crisis in the early 1930s would bring about a significant reduction in migration (Bade 2001).

Restrictive immigration policies, economic problems, and economic crisis brought about a total re-articulation of migration processes and prevented the restart of the mechanism of labor transfer that had characterized the entire first globalization. Italian politicians pointed to a resumption in emigration, which seemed to take place immediately after the war with over 600,000 expatriates in 1920. In the following years, though the values decreased, they remained at high levels (between 141,000 and 390,000 units up to 1931), but they did not reach the size recorded before the war. Fascism pursued, initially, the same policy followed by liberal governments of promoting emigration. It was the "speech of the Ascension" of 26th May 1927 that initiated a change in direction (Nobile 1974). The regime introduced increasingly stringent regulations that added an internal obstacle to the external ones already posed by the immigration countries and by the difficult economic conditions (Nobile 1974). The overall result was a further sharp reduction in expatriates that, since 1932, went down to below 100,000 units annually.

From the statistical point of view, the collection of data on international migration experienced several changes and improvements despite the great difficulties the official statistics encountered until the creation of Istat in 1926. In 1914 changes in the definition of migrant were introduced to take into account the new definition of the law aiming at the legal protection of migrants of 1913. The source, therefore, regarded those who went abroad to perform manual work, to operate small businesses, or to reach family members who had already emigrated for work reasons. From 1928 onwards these groups were joined by the intellectual workers. From 1915 the summary models compiled by the prefectures were replaced by individual records processed directly by the DGS.

Since 1921 the survey of returnees was also extended to countries in Europe and the Mediterranean. But the most significant change, also introduced in 1921, for the statistics of the phenomenon, was the use of collecting coupons included in the passports and withdrawn when boarding, disembarking, or crossing the border. The information thus collected was supplemented for flows to and from non-European countries by the lists of names of those on board. This innovation allowed for the overcoming of some of the shortcomings of the survey described above. Critical areas that remained were related to illegal migration, the multiple departures throughout the year, the use of passports granted for reasons other than emigration, and the ineffectiveness of border controls. According to CGE (1926) these problems concerned mainly the flows to European countries. The introduction of these new methods coincided with the passage of the responsibility of surveying to the CGE. With the removal of the CGE, in April 1927, the office responsible for surveying passed to the Directorate-General for Italians Abroad and, since November 1929, to the ISTAT.

4 The European Labor Migration Period

The end of World War II marked the opening of a new phase in the European migration scenario. In fact, after the end of the first post-conflict emergency most of the Western European countries experienced three decades of extraordinary economic growth and in which immigration played a major role. As for Italy, the end of hostility and the fall of fascism marked a return to a policy of active encouragement of emigration. The country aimed to resume the same role it had had during the first globalization and the right to emigrate was present in the Constitution of the Republic, as if to confirm a clean break with the Fascist policy of autarkic closure. The political orientation in favor of emigration resulted in two different levels of intervention in a situation characterized by a strong commitment by the entire governmental structure and the public apparatus to achieve the objective of maximizing the output fluxes (Bonifazi 2005). On the one hand, Italy aimed at concluding bilateral agreements with countries willing to accommodate incoming flows; on the other, it sought to promote the Italian interest of encouraging emigration internationally, particularly concentrating the efforts within the international organizations that saw the light in those years.

The output volume grew rapidly, returning in 1947 to 254,000 units. Until near the end of the 1970s, the number of expatriates, even reflecting changes dependant on the phases of European labor markets, kept the numbers up to between 200,000 and 387,000 units. In the first part of this period, Italian workers formed the bulk of European migration for work purposes. This resurgence in migration coincided with a period of extraordinary growth in the Italian economy and especially the final transformation in the industrial sense of the production structure of the country. The result was that, for the first time since unification, interregional internal migration came to be seen as a real alternative to emigration towards foreign countries (Sonnino 1995). In fact, between 1955 and 1965, the South and the North East lost 1.7 million inhabitants in the interchange of internal migration, while at the same time the total loss of the whole country to foreign countries was just under 1.5 million.

Since 1955 data of population registers is also available on cancellations and registrations to and from other countries, which represented a new source for the measurement of international migration. In reality, however, the source of reference remained the survey of expatriates and returnees for the whole period considered. In this source the definition of emigrant was further enlarged in 1943 to include those who went abroad to pursue a profession, art or craft, or just under the dependency of others to follow or to join family members expatriated for such reasons and finally for those who for whatever reason wished to establish residence outside national boundaries. The coupon system was abandoned in 1955 and replaced, for the flows to European countries, by verifications made by the Italian municipality of residence (or former habitual residence) of the migrants. Since 1969 these criteria were extended to the movement towards non-European countries where, since 1955, the boarding lists and the reports of expatriates registered via airplane were used.

With these changes, which led to the creation of files of immigrants and emigrants bound for foreign countries and held by municipalities (Bonarini 1976), ISTAT tried to adapt the survey to the great changes that had characterized the global migration scene in the meantime, especially in the European context. In fact, many of the assumptions and requirements on which the survey was based were actually exceeded. In particular, the birth of the European Community had liberalized much of the movement and migrating could be done without leaving the necessary track for the proper functioning of the statistical survey. The result was a progressive loss of information capacity from a source that had accompanied a century of Italian history and would be completely abandoned in the 1980s.

5 The Mass Immigration of the Second Globalization

The oil crises of the early 1970s marked the end of the golden period of European Labour Migration. In the Italian case, it determined the prevalence of returns on the departures and the closure of a migration cycle that had opened before the unification of the country. The first flow of foreign immigrants that started towards the end of the decade faced a substantial shortage in legislation and an equally substantial information gap (Bonifazi 2007). The available statistics were limited to census data on foreign residents and some information on residence permits granted by the Ministry of the Interior, while the population registers data, which noted the in and out movements of foreigners, gave a total value and did not distinguish the foreign residents from the Italians. The radical change in the dynamics of migration posed the need to redirect a statistical system which captured outflows, then beginning to decline, yet that was unable to give an account of the incoming movements. On these grounds, in the early 1980s, the scientific community began to play an important role in stimulating, encouraging, and proposing solutions that would enable the national statistical system to provide information on a phenomenon still in its early phase (Natale 1983).

Meanwhile, the political interest in the phenomenon grew. In 1986 the first law on immigration was passed and the debate became more and more lively. As had happened a century before the migration, even in this case, the debate between supporters and opponents found the size of the phenomenon as the first natural terrain for confrontation. A war of figures began, supplied by the little and controversial data available. The census of 1991 sought to improve the quality of data collected on the phenomenon. A new survey on foreigners registered in the municipal registry saw the light and a satisfactory form was given to the statistics on the permits to stay, eliminating the problem of duplication and missed cancellations. Furthermore information on foreigners was included in many current surveys.

Meanwhile, foreign immigration was consolidating its position within Italian society, becoming a structural element of the reality of the country. The foreign residents moved from 211,000 units in 1981 to 356,000 in 1991. Due to the fall of the Berlin Wall in 1989 and the consequent end of the socialist regimes, foreign immigration in Italy marked the beginning of a growth phase that, over the years,

became tumultuous: in 2001 foreign residents surveyed were 1.3 million and 10 years later they would be 4.5 million recorded in the municipal registers.

In the 1990s immigration gained weight in the Italian political debate. Overall, however, the focus of politics towards statistical information on the phenomenon has been limited and sporadic. The impression is that in recent years, the national statistical system acted in a substantially independent manner in trying to improve and expand its ability to collect data on the size and characteristics of the phenomenon.

In recent years, the final settlement of the surveys on permits to stay and aliens entered into the population register took place. These offered a good base for information on the phenomenon and the census of 2001 gave the foreign presence the attention it deserves. More recently, the availability of the data on foreigners collected in the labor force survey and the distribution by sex and age of the foreign population residing in municipalities have recovered part of the delay we had when compared to other EU countries. Much remains to be done, particularly in terms of timeliness, because the knowledge demand is increasing and is becoming more pressing over time due to the growth in the scale of the phenomenon.

6 Conclusions

The 150-year period since the unification of Italy has seen extraordinary changes in migration statistics. These changes are related to the evolution of the phenomenon and of political needs. From this last point of view, it should be emphasized that the statistical knowledge of migration is an integral and decisive element of the decision-making process (Kritz 1987). This function, which reflects a modern understanding of the relationship between decision making and statistical information, was realized by the Italian statistical system in different ways during the long period of time considered. Certainly, statistics on emigration have been an element of reference in political discussions during the Liberal Italy, also thanks to the actions of an important figure, such as Luigi Bodio. A more instrumental relationship developed during fascism, when a clear contradiction between the undoubted improvements in the statistical system and on specific surveys, and an interest in pursuing decidedly authoritarian regime objectives in the field of migration, opened.

Paradoxically, the political role of statistical information on migration appears to have declined in recent years. In particular, when considering foreign immigration, the degree of integration and interaction between politics and information has appeared decidedly modest. A central role that statistical information should have in the decision-making process has not been recognized. Often the task of remedying the lack of clear political decisions of policy and organization of the overall system of data collection has been left to the good will of individuals or individual agencies that deal with the problem (Bonifazi and Strozza 2008). The necessary clarity on an essential element of political debate has frequently been lacking and has often fueled a war of numbers that has certainly not contributed to the serenity of the debate. In the meantime new challenges are entering the agenda of the statistical

system (Bonifazi and Strozza 2008). The stabilization of foreign immigration and the growth of a second generation born or brought up in Italy are posing the need to collect information on naturalized foreigners and on the population of foreign origin. Furthermore integration issues are gaining a central role in the evaluation of the overall impact of immigration on Italian society. Some of these questions have already been addressed by official statistics and all are at the core of scientific debate. The long journey of Italian migration statistics is far from concluded, it has a long future in front of it.

References

Bade, K.J.: L'Europa in movimento. Le migrazioni dal settecento ad oggi. La Terza, Bari (2001)
Bodio, L.: Intervento alla seduta del 26 marzo 1877. In: Ministero di Agrioltura, Industria e Commercio – Direzione di Statistica Ann. di Stat. Serie 1ª, vol. 88, pp. 160–169. Tipografia eredi Botta, Roma (1877)
Bonarini, F.: Analisi della rilevazione del movimento migratorio con l'estero. Genus **32**(1–2), 141–177 (1976)
Bonifazi, C.: Dall'emigrazione assistita alla gestione dell'immigrazione: le politiche migratorie nell'Italia repubblicana dai vecchi ai nuovi scenari del fenomeno. Pop. e Storia **1**, 19–44 (2005)
Bonifazi, C.: L'immigrazione straniera in Italia, 2nd edn. Il Mulino, Bologna (2007)
Bonifazi, C., Strozza, S.: Informazione statistica ed esigenze conoscitive sull'immigrazione straniera: realtà, problemi e prospettive. In: Istat, La presenza straniera in Italia: l'accertamento e l'analisi. Atti del Convegno Roma, 15–16 dicembre 2005, pp. 187–212. Istat, Roma (2008)
Commissariato Generale dell'Emigrazione: Annuario statistico della emigrazione italiana dal 1876 al 1925. Edizioni del CGE, Roma (1926)
Congrès général de statistique: Compte rendu des travaux du Congrès général de statistique réuni a Bruxelles les 19–22 septembre 1853. M. Hayez, Bruxelles (1853)
Kritz, M.M.: International migration policies: conceptual problems. Int. Migr. Rev. **21**(4), 947–964 (1987)
MAIC – Direzione di Statistica: Ann. di Stat.. Serie 1ª, vol. 6, 1875 Sec. edn. Tipografia eredi Botta, Roma (1880)
Marucco, D.: L'amministrazione della statistica nell'Italia unita. Laterza, Bari (1996)
Marucco, D.: Le statistiche dell'emigrazione italiana. In: Bevilacqua, P., De Clementi, A., Franzina, E. (A cura di) Storia dell'emigrazione italiana. Partenze, pp. 61–75. Donzelli, Roma (2001)
Ministero di Agricoltura, Industria e Commercio: Ann. 1872 Parte II – Statistica. Tipografia Sacchetto, Padova (1872)
Natale, M.: Fonti e metodi di rilevazione della popolazione straniera in Italia. Studi Emigrazione **20**(71), 265–296 (1983)
Nobile, A.: Politica migratoria e vicende dell'emigrazione durante il fascismo. Il Ponte **30**(11–12), 1322–1341 (1974)
Ostuni, M.R.: Leggi e politiche di governo nell'Italia liberale e fascista. In: Bevilacqua, P., De Clementi, A. e Franzina, E. (a cura di) Storia dell'emigrazione italiana. Partenze, pp. 309–19. Donzelli, Roma (2001)
Sonnino, E.: La popolazione italiana dall'espansione al contenimento. In: Storia dell'Italia repubblicana: vol. 2, Tomo I. La trasformazione dell'Italia sviluppo e squilibri, pp. 532–585. Einaudi, Torino (1995)
Sori, E.: L'emigrazione italiana dall'Unità alla seconda guerra mondiale. Il Mulino, Bologna (1979)
Statistica Generale del Regno d'Italia: Censimento degli Italiani all'estero. Stamperia Reale, Roma (1874)

The Evolution of Statistic Information on Agricultural Labour Force Through Italian Agricultural Censuses from 1961 to 2010

Loredana De Gaetano

Abstract

The main purpose of this paper is a brief analysis on the trend of the collection of statistical information on the holdings' labour force in six Italian agricultural censuses (1961–2010). The aim is to show how the agricultural holding's management has actually changed in terms of entrepreneurship supported by a sufficient level of agricultural expertise and, especially, to identify and quantify the recourse to other workforce than strictly family by the agricultural holdings. The paper illustrates the evolution from 1961 to 2010 occurred in the information of Labour Force, included in a specific Section in the holding's questionnaire of each census, in the light to offer as possible to all the users a comprehensive picture of the changes of agricultural working reality in Italy.

1 Introduction

Unlike what is commonly believed, the agricultural work has not only a very significant component of seasonal or occasional type related to the productive seasonality, but also a significant rate of permanent and fixed-term, very structural, employment. In recent years there has been an evolution of the professionals involved in the holdings labour, corresponding to changes and productive differences in the agricultural holdings for which statistical knowledge is also required for the adoption of appropriate national and Community legislation (Regulation (EC) No. 1166/2008). Beyond with the traditional farming and livestock activities, the farmers started to devote more and more not agricultural activities but however connected with agriculture, such as agro-tourism, processing, manufacturing and marketing of

L. De Gaetano (✉)
Istat, Rome, Italy
e-mail: degaetan@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_7,

products, third partly work in other holdings, protection and preservation of land, energy production. As part of the agricultural employed labour force, the presence of non-EU workers has a more significant importance, which currently is about 13 % of the total labour force (Cicerchia and Pallara 2009; Inea 2009).

It is clear that the specialization of production, the competitiveness of markets and the multi-functionality of holdings have required also dynamic and rapid changes in the types of agricultural labour force involved in the agricultural activities in order to adapt them to the market demand. Although the presence of the direct work of the holder and his/her family has been confirmed, the workload in the holding is shifting more and more on contractual employees.

The analysis of the results and changes in the composition of the agricultural labour force in the farm structure surveys, with particular reference to the censuses from 1961 to 2010, clearly shows the changes in entrepreneurship in agriculture, making it possible to know the main types of holdings that characterize the different realities actually present in Italian agriculture.

The knowledge of these types of holdings is of particular importance since the ability to deal with rapid market adjustments related to new and different economic position that the agricultural sector is taking on in society in rapid transformation, as well as the increasing integration of countries of European Union, depends on them.

However, it shoud be pointed out that the acquisition of information on agricultural work in general is strongly influenced by international and national legislative constraints, which in statistical surveys end up with imposing methodological choices on the types of information to investigate. So, like any other holding's characteristics (crops grown, livestock, machinery, methods of production, sales of holding production, rural building, etc.), even the agricultural work, broadly speaking, and the categories of farm workers, are strongly bounded by the rules of EU and FAO recommendations.

In fact, in the course of time, specific regulations and directives of the EU Commission and Council have been promulgated for each census (and related intermediate sample surveys), making compulsory questionnaires or list of variables to be collected by Member States.

In addition it's also the national constraints, arising from the happened and ongoing transformations in society in general as well as in national legislation relating to agricultural contracts, types/subjects authorized to carry out the agricultural activities.

2 The Evolution of the Needs of Information

Bounding the analysis to agricultural work and the types of farm workers, in each of the census questionnaires from time to time used, a specific section on farm labour has been dedicated, in which essentially questions aimed at detecting the single categories of farm workers, with a minimum set of variables of socio-demographic

(sex, date of birth and even professional status) and economic (working time spent in the holding, part-time and multi-activity).

In order to better understand the level of the management of holding, some questions were aimed at assessing the level of agricultural professionalism of the manager through the educational level acquired (degree or certificate in agricultural science, intermediate vocational schools, primary school or no education).

The framework in which to insert, from time to time, the mentioned general variables with the appropriate modifications and/or supplements has been articulated as follows:

(a) Family labour force
(b) Other non-family labour force
(c) Other gainful activities (agricultural and non) of the holder and his/her family
(d) Educational agricultural level of the holder/manager

It has been understood that the basic definitions (agricultural work, working days, macro-categories of labour force, professional status, other gainful activity, holder, manager) have remained unchanged for all six Censuses carried out until now, and the changes taken into account in the Italian censuses for the factor "Work" have been essentially of two types:

(a) *Basic*: the age of family labour force to be considered for the agricultural works has been increased from 14 years (the first two censuses) to 15 (1982 and 1990) and to 16 years (2000 and 2010), on the basis of raising the minimum age for compulsory education;

(b) *Specific*: from time to time the type of information collected reflects the information needs at Community level and those expressed in the national and local level. This affects the structure of the part (section) of the questionnaire of the holding used.

(1) ***Census 1961***: in the specific Section VII (Labour force) inserted in the questionnaire of the holding the labour force involved in the holding at the date of the census (April 15, 1961) had to be indicated, split by sex, age and category, even if temporarily absent due to illness, vacation, etc. People occasionally occupied at the census date (labourers, daily workers, etc.) were excluded.

    In an another question, dedicated to some special news on the activities of the holder and his/her family, it was asked if the holder and one or more of his/her family members who worked exclusively or mainly in the holding carried out gainful activities, agricultural or non-agricultural, even in other holdings. It was also asked if other family members of the holder carried out occasional activities in the holding.

    The section ended with an indication of the number of working days normally performed in holding by the farm labourers, daily workers and similar during the agrarian year, split by sex, as well as with an indication of the total area of land of holding managed by "share partners" and assimilated (Istat 1961).

(2) ***Census 1970***: also on this occasion, in the questionnaire of holding a specific Section IV (Labour) was designed, divided into two separate parts: one dedicated to the family labour (holder, spouse and other relatives of the holder)

and the second one to the non-family members (officers and employees, daily or occasional workers, and similar, improper settlers and similar).

This section ended with the question on the title of study of the holder and with information on other gainful activities carried out by the holder and his/her family (Istat 1970a, b).

(3) ***Census 1982***: The content of the section of the questionnaire on the farm labour force (Section VIII) was substantially unchanged by comparison with the analogous section of the previous census. The information on the farm work was specified better with the distinction of the labour force aged 14 and over who had worked in the holding, further divided in according to the familiar categories (holder, holder's spouse and other relatives of the holder) and non-family labour force. In 1982, in order to better assess the changed contractual relations, the part referred to other labour force than family members has been further divided into
   – Not fixed-term workers, special categories, employees and managers (for each of them the sex and the number of working days during the agrarian year 1981–1982 were recorded).
   – Fixed-term workers, improper settlers and similar, for which, as in the past, was only required the total number of persons by sex.

   In addition, for each member of the family labour force, other than the holder, the traditional information on their gender, year of birth and annual number of working days have been integrated with the indication of any gainful activity carried out (as exclusive, main or secondary activity using specific code of the economic sector where such supplementary activities was carried out) (Istat 1982a, b).

(4) ***Census 1990***: as already anticipated and unlike in the past, with the Section V (Labour) of the questionnaire information on all members of the household of the holder, present in the holding in the agricultural year of reference were required. In practice, this time the information have been extended to the entire household of the holder, understood in demographic sense, including its members who were not working in the holding. Moreover, among the categories of family labour force the relatives of the holder who lent their work in the holding were entered separately. The information on the work also focused on non-family labour force, with the distinction between (Istat 1990):
   (a) Permanent workers, special categories, employees and managers, each of which was required by sex, year of birth and the number of working days carried out during the agricultural year of reference;
   (b) Fixed-term workers, in terms of number of working days divided between males and females;
   (c) Improper settlers and assimilated, also in terms of number of working days divided between males and females.

(5) ***Census 2000***: How for the past censuses a specific Section of the questionnaire (Section VII) has been dedicated to the labour. Its structure is quite similar to that of the previous census. Information on the number of working days carried

out during the agrarian year of reference continued to be taken into account, with reference to the agricultural labour force aged 16 and over.

In detail, the categories taken into account for the holder's family and relatives were the holder; the holder's spouse; other family members, separately for those who worked in the holding and for those not working; the relatives of the holder who worked in holding. For each of them it was compulsory to indicate, as in 1990, sex, year of birth, employment status held in the week preceding the census reference date, working days carried out during the agricultural year 1999–2000, and other gainful activities out of holding, specifying whether that activity was exclusive, main or secondary, as well as the economic sector where the activity was mainly carried out and position in this gainful activity. The information on the non-family labour force were required with respect to directors, employees, workers and assimilated categories, separately for the not fixed-term employees and fixed-term workers. For the first category it was required sex, year of birth and working days performed in the holding in the agrarian year, while for fixed-term labour force it was to specify, separately for males and females, the total number of people employed and the relative number of working days performed in the agrarian year (Istat 2000).

Also in 2000 as in the past the following information have been requested: manager by type (holder, spouse, etc.), sex, year of birth, the number of working days performed during the agrarian year, the title of study acquired (whether in agricultural schools and/or other types of schools, with a detail of the official titles of study in accordance with the law school).

Similarly to 2000, it was asked if the manager attended professional courses lasting not less than 3 months, with the issuance of a certificate, and if he/she dedicated to the improvement of the professional ability in the agricultural sector.

(6) ***Census 2010***: the structure of family labour force remains the same. A change occurs in collecting information on the time worked: besides the number of working days performed in the holding, also the average daily duration in hours has been required.

Similarly, the question relative to the manager remains more or less unchanged. However, the structure and type of information on the non-family labour force has been renovated. The traditional schematization of the categories of other workers divided into those permanent and temporary (managers, employees, workers, etc.) was substituted by the distinction of other workers in "permanent or continuos way" and in "occasional" form. The first macro-category includes "the people who in the agrarian year 2009–2010 have worked continuously in the holding interviewed, regardless of the weekly contract" and the second "people who in the agrarian year 2009–2010 have not worked continuously in the holding, such as short-term jobs to perform seasonal work or single specific work phases". For each component of the labour force in continuous form, information to citizenship (Italian, other EU country, non-EU) and the percentage of working time spent on holding but dedicated to non-agricultural activities related to the holding (Istat 2010) have

been collected in addition to the sex, the year of birth and the working time (number of days worked and average hours per day).

For fixed-term workers, however, it has been recorded only the total number of males and females broken down by nationality (Italian, other EU country, non-EU), sex and the number of working days performed in the holding for the agricultural activities and for those related activities.

Then, to better understand the workload for production activities it has also been included in the questionnaire another question dedicated to the total number of "workers not employed directly by the holding (e.g. contractors' employees) carrying out agricultural or related activities". For this type of workers it has been requested the number of persons according to the Citizenship (Italian, other EU country, non-EU) and the number of working days performed in the holding for carrying out the agricultural and related activities converted into 8-h days (Istat 2010).

3 The Offer of Information: The Statistical Context of Reference

3.1 Generality

After the first experience of the 1961 census, the publication of the results of the Census of Agriculture 1970 provided the basic subject for analyses and debates of researchers and experts achieving results always more significant in understanding different types of holdings. The interpretative models initially adopted have been based essentially on an approach "dual" based on the contrast between capitalist and peasant holdings. These analyses have proved to be effective for a correct interpretation of the reality of structural and social components of Italian agriculture in the 1970s, although the onset and the roots of new phenomena have progressively put into question the validity of this approach (Barberis and Siesto 1974; Brusco 1979; Calza Bini 1976; Cosentino et al. 1977; Pugliese and Mottura 1975). During the 1980s, with the spread of a new model of Italian economic development, focused on small and medium-sized enterprises and the spread of industrialization in many regions, the analysis was designed to understand and interpret the changes in agricultural structures, favouring, on the one hand the part-time agriculture, understood as a new holding structural and management realty of the holdings, not temporary, and on the other hand the growing internal structure of family holdings. This has strengthened the phenomenon of "persistence" of the family holdings in which the internal articulation and differentiation of the main types of holdings were attributed to a different role from the point of view of production, employment or social depending on the territorial context where they were located. At the same time, the analyses of structures and entrepreneurship inside the Italian agriculture were dedicated to capture and interpret the processes of outsourcing of holdings with more intense use of services by third parties (subcontracting), as well

as the processes of IT innovation and integration with the surrounding territory, considering the farm as well as "open and enlarged entity" (Fanfani 1989).

Just over 80 years the family holdings show the greatest changes, with a sharper differentiation of the internal forms of family that ratifies definitively the achievement of part-time and family multi-activity as phenomena that are not temporary but permanent of the Italian agriculture. To this it is to be added a significant aging of the holders, especially in small holdings, together with an increased use of services from contract enterprises by third for small holdings, especially for farming operations that require the use of mechanical means economically costly. These changes occur concurrently with a general tendency towards specialization and concentration of production in a limited number of holdings. A brief analysis of these substantial changes may therefore provide a new overview of the structure of holdings, in its dynamic evolution. The data needed to understand all these changes are provided by the general agricultural censuses carried out from 1961 to 2000.

3.2 The Agricultural Holding According the Type of Management

The results of agricultural censuses up to 2000 clearly indicate that Italian agriculture is characterized by a massive presence of a family holdings, in which responsibility for production activities is in the hands of a holder who participates manually in the agricultural work of the holding. In fact, the rate of holdings under a direct management in 1961 was 81.2 % and in 2000 the incidence rose to 94.8 %. In this regard, however, it is to point out that in 1961 sharecropping was still widespread sharecropping (7.4 % of surveyed holdings with 11.0 % of total area land); over time this type of management has been disappearing, after its abolition by law in 1973, so that in 2000 its life was greatly reduced to 0.1 % in terms of holdings interested and total area. In detail, the direct management has a decline of 10.5 % in the period 1961/1970 which is matched, however, to an increase of 8.7 % in terms of total area land. In the next period 1982/1970 the decrease of such holdings is strengthened with a further slight decrease of 1.9 % while continuing to further increase the total area of 11.4 %.

3.3 The Size of the Holding

The distribution of holdings and its total area land by size shows how in the agricultural sector in the period 1961–2000 the presence of micro-holdings or holdings in which the total area covers a small fraction of the total area in 1961 is always very massive.

Indeed if from Agricultural census 1961 the holdings with less than 5 ha of total area land (excluding those with no area land) were 75.2 % of the whole surveyed, having, however, just 1/5 of the national area land, from the results of census 2000 the rate of these types of holdings, rather than decreased following all the actions

Table 1 Number of holdings and relevant number of working days by categories of labour force

	Year of census							
	1970		1982		1990		2000	
Categories of labour force	Holdings	Working days	Holdings	Working days	Holdings	Working days	Holdings	Working days
Holder	3,607,298	365,928,651	3,248,968	289,125,283	3,002,127	216,575,877	2,578,794	175,571,828
Spouse	–	–	1,286,976	108,117,809	1,400,437	76,446,256	1,087,347	53,653,405
Family members and relatives	1,958,305	319,387,820	822,199	110,318,228	860,120	88,304,542	623,045	54,830,569
Permanent non-family workers	103,739	48,421,956	61,035	28,151,368	38,216	17,277,663	34,551	13,335,558
Temporary non-family workers	1,355,919	125,724,193	660,544	73,104,657	621,153	61,922,122	378,737	36,156,468
Total	3,607,298	859,462,620	3,269,170	608,817,345	3,023,344	460,526,460	2,594,815	333,547,828

Source: Istat

and policies related to land consolidation, abandonment of small or marginal, etc., increased to 77.7 %, having, however, just 14.5 % of the total area land surveyed.

Still more particularly, the larger holdings (100 ha and over) which in the period up to 1982 had an increase in the number of holdings and its total area, according to the results of the 1990 census show a downward momentum, recording decreases in the period 2000/1990 equal to 8.3 % in terms of holdings and 11.2 % in terms of total area land showing as total balance of the 40 years between 1961 and 2000 more limited decreases of 2.2 for holdings and 4.1 % for the total area land.

3.4 The Labour Force

The transformations of the types of management have been accompanied by coherent changes, of the working days performed in the holding. The total number of annual working days decreased from 859 million in 1961 to about 334 million in 2000 (Table 1), with a decrease of more than 61 % compared to 1970. Among the family labour force the holder reduces its working engagement of 52.0 % compared with a corresponding decrease of family holdings equal to −28.5 %. Much more also the working engagement of his family, including spouse and any cooperator relatives decreases with a rate amounting to −82.8 % in terms of number of working days and to −68.2 % for the number of persons. As regards the temporary and permanent workers a significant reduction resulted in terms either of number of holdings engaging such categories or of number of working days performed in the holding (respectively −67 and −72 %).

References

Barberis, C., Siesto, V.: Produzione agricola e strati sociali. F. Angeli (1974)
Brusco, S.: Agricoltura ricca e classi sociali. Milano, Feltrinelli (1979)
Cicerchia, M., Pallara, P.: Gli immigrati nell'agricoltura italiana. INEA, Roma (2009)
Calza Bini, P.: Economia periferica e classi sociali. Editore Liguori (1976)
Cosentino, V., Fanfani, R., Gorgoni, M.: Alcuni aspetti dello sviluppo dell'agricoltura meridionale dal dopoguerra ad oggi. In: Investimenti e disoccupazione nel Mezzogiorno (a cura di Augusto Graziani ed Enrico Pugliese) (1977)
Fanfani, R.: Il contoterzismo nell'agricoltura italiana. INEA, Il Mulino Ed., Bologna (1989)
Inea, Il lavoro. In: Annuario dell'agricoltura italiana. Edizioni Scientifiche Italiane, Napoli (annate varie) (2009)
Istat, 1° Censimento generale dell'agricoltura – 15 aprile 1961. Relazione preliminare. Roma (1960)
Istat, 2° Censimento generale dell'agricoltura e Rilevazione dei dati per la istituzione del catasto viticolo – 25 ottobre 1970. Disposizioni e istruzioni per gli Organi periferici. Roma (1970)
Istat, 2° Censimento generale dell'agricoltura – 25 ottobre 1970. Volume VII – Atti del censimento. Roma (1975)
Istat, 3° Censimento generale dell'agricoltura – 24 ottobre 1982. Disposizioni e istruzioni per gli Organi periferici. Roma (1982)
Istat, 3° Censimento generale dell'agricoltura – 24 ottobre 1982. Relazione generale sul censimento, Volume IV. Roma (1991)

Istat, 4° Censimento generale dell'agricoltura – 21 ottobre 1990. Istruzioni per la rilevazione dei dati. Roma (1990)
Istat, 5° Censimento generale dell'agricoltura – 22 ottobre 2000. Istruzioni per la rilevazione. Roma (2000)
Istat, 6° Censimento generale dell'agricoltura 2010. Istruzioni per la rilevazione. Roma (2010)
Pugliese, E., Mottura, G.: Agricoltura, mezzogiorno e mercato del lavoro. Il Mulino, Bologna (1975)
Regulation (EC) No. 1166/2008 of the European Parliament and of the Council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation (EEC) No. 571/88

Changes in the Geographical Distribution of Inhabitants in Tuscany Since 1861

Luca Faustini, Linda Porciani, Graziella Sanna, Cristiano Tessitore, and Alessandro Valentini

Abstract

Geographical and administrative approaches provide a framework able to explain overtime variation in inhabitants distribution. In this chapter Tuscany Census Data are employed from 1861 (the unification of Italy) to 2001 (the most recent available census). A mathematical procedure able to remove the effect of (time) changes in local boundaries has been applied in order to compare Tuscany Municipalities over time and to better highlight modifications in the residential profile of the whole Region. Classification of municipalities by "crowns" represents a special focus of the analysis.

1 Background

In 1861 around 1.9 millions of inhabitants lived in Tuscany. In 2001 this number has nearly doubled (3.6 millions). However, variations were not homogenous in the whole Region: for instance, during the last 150 years the population of the Municipality of Florence (Regional Capital) grew about 400 %; in the same period the number of inhabitants in the whole Province grew of around 100 %. On the other hand the Province of Siena remained somewhat unchanged.

Despite this paper is the result of the work of all the authors, the paragraph sub-division is the following: Luca Faustini (Sects. 5.2 and 5.3), Linda Porciani (Sects. 1 and 2), Graziella Sanna (Sect. 6), Cristiano Tessitore (Sects. 4 and 5.1), Alessandro Valentini (Sect. 3).

L. Faustini (✉) • L. Porciani • G. Sanna • C. Tessitore • A. Valentini
Istat – Regional Office for Tuscany and Umbria, Lungarno Colombom, 54 - 50136 Florence, Italy
e-mail: faustini@istat.it; porciani@istat.it; grsanna@istat.it; tessitore@istat.it; alvalent@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_8,

In order to highlight the role that geographical features (altimetry, proximity to the coast, and so on) play in affecting the settlement profile of Tuscany population, intra-period municipalities legislative boundaries changes have been removed (Faustini et al. 2011). The analysis has been then performed following three steps: (1) removing perturbations due to transformations in the administrative boundaries (via mathematical procedures); (2) classifying municipalities according to some significant criteria; (3) linking each group of towns to their population changes.

2 Data

The population census is an effective source for the observation of society and a unique way to analyze changes in the medium and long-term period. In this study, we focused our attention on Tuscany Census Data beginning from the first Italian census (1861) to the last available one (2001).

As previously mentioned, it has been necessary to remove local boundaries modification over time to make data comparable.

In 150 years some municipalities disappeared, some born, and many others experienced annexation and/or splitting-off of their confines. The comparison has been realized using a method, explained in Sect. 3, in order to actualize the local borders.

The territorial domain of analysis is represented by various types of aggregations: municipalities and Provinces[1] as administrative domains, altimetry/proximity to the coast and crown as geographical domains.[2]

3 Methodology

In 1861 the geographical territory of Tuscany was administratively divided into 269 municipalities (Istat 2001, 1994). Despite borders of the Region haven't been varied over time,[3] legislative actions changed the boundaries of nearly half of

[1]Tuscany is actually divided in ten provinces: Massa Carrara, Lucca, Pistoia, Firenze, Prato, Livorno, Pisa, Arezzo, Siena, Grosseto.

[2]Data used in this chapter are available on a Web Information System, called SITO. DEM (Porciani et al. 2011), containing basic census data since 1861–2001 and in a more detailed way data since 1951–2001. All of them are disaggregated by municipality and by other various spatial aggregations (such as Health Areas and Local Labor District). The database moves from Health for All (HFA) data base, which is a project promoted by WHO, and follows the same data structure (WHO Regional Office for Europe XXXX).

[3]The only exception is represented by the acquisition of about 200 residents from Fiumalbo (MO) in the 1950.

the municipalities and of various Provinces (Prato, for instance, was separated by Florence less than 20 years ago) affecting around 63 % of the total population. More in detail 40 municipalities have been created; 22 was dissolved; other municipalities experienced only partial increases or reductions (or both) of their territories. As a consequence of these changes between 1861 and 2001, it is firstly necessary to estimate the total population amount for each municipality at every preceding censuses considering actual boundaries.

Assuming $P(i, t)$ to be the population (at the borders of the time) of municipality i at census t; $\hat{P}(i, t)$ to be the estimated level of such municipality with actual borders at time t, the following procedure was used:

- If during the period 1861–2001 no changes occurred in the borders, than $\hat{P}(i, t) = P(i, t)\ \forall\, t$;
- If at time t^* a legislative change in the borders implies a shift of X persons from i to j, $X(i, j, t^*)$, the (historical) estimate of both populations at actual boundaries is:

Municipality i (posting of boundaries)

$$\hat{P}(i,t) = \begin{cases} P(i,t) & \text{if } t > t^* \\ P(i,t)\left[1 - \left(\frac{X(i,j,t^*)}{P(i,t^*)}\right)\right] & \text{if } t \le t^* \end{cases}$$

Municipality j (annexations)

$$\hat{P}(i,t) = \begin{cases} P(j,t) & \text{if } t > t^* \\ P(j,t) + P(i,t)\left[1 - \left(\frac{X(i,j,t^*)}{P(i,t^*)}\right)\right] & \text{if } t \le t^* \end{cases}$$

More complex cases (multiple annexations in subsequent years, annexations and detachments, and so on) are treated in a similar way.

4 Municipality Classification

Municipalities were classified according to various features. From an administrative point of view, we used two different splits: Provinces and "crowns." While the first criterion is widely applied in population studies, the second one is quite innovative. In particular, given a Province, three different "crowns" are identified (see Fig. 1 for an example):

- Provincial Capitals → crown 0;
- Adjacent to Capitals (neighbors of provincial Capitals) → crown 1;
- Other municipalities → crown 2.

From the geographical point of view, we focused on a mixed criterion based on altitude and proximity to the coast (Altimetry Zone: henceforth ZALT). Municipalities are then classified in:

- Inland mountain [IM],
- Coastal mountain [CM],
- Inland hills [IH],
- Coastal hills [CH],
- Plains [P].

A cross-distribution of the municipalities by ZALT and crown is provided in Table 1. From the altitude point of view, Tuscany Municipalities are settled mainly

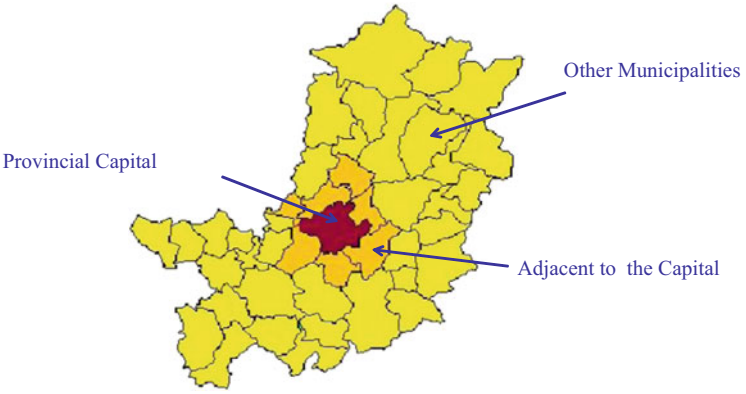

Fig. 1 Province of Florence: crowns classification (boundaries 2001)

Table 1 Municipality distribution by crowns and ZALT

Crown	ZALT					
	Inland mountain	Coastal mountain	Inland hills	Coastal hills	Plains	Total
Provincial capital	1	1	4	1	3	10
Adjacent to the capital	9	2	30	6	5	52
Other municipalities	66	2	106	34	17	225
Total	76	5	140	41	25	287

in Inland Hills (140 cities), Inland Mountains (76), Coastal Hills (41), and Plains (25). Only five municipalities are located in Coastal Mountains: considering their relative size, there is not enough evidence to highlight any kind of detailed trend.

From the crown point of view, the vast majority of municipalities are located in other municipalities (225), and the rest is divided in Adjacent to the Capital (52) and Provincial Capital (10). The same distribution affects every altitude area except for Coastal Mountain where, again, the distribution is affected by a small amount of municipalities.

For each of the above classifications, the average annual rate of population change between 1861 and 2001 (equal to 4.3 % for the whole Region) is illustrated in Table 2. In terms of administrative point of view, Prato is the most dynamic Province of the Region (+10.2 %), with a population incidence nearly of 6.5 %.

In terms of crowns, the Provincial Capitals (Crown 0) average annual rate of growth is 5.7 %, while for other municipalities (Crown 2) is 3.1 %.

Considering the geographical classification (ZALT), the growth rate is more intensive in Coastal Mountain (+8.2 %) and Plains (+5.6 %).

Table 2 Average annual rate of change (per 1,000) between 1861 and 2001 for various groups of municipalities (in parenthesis the percentage weight on the regional population in 2001)

Provinces	
Massa Carrara	+4.8 (5.7 %)
Lucca	+3.0 (10.6 %)
Pistoia	+4.5 (7.7 %)
Firenze	+4.9 (26.7 %)
Prato	+10.2 (6.5 %)
Livorno	+5.3 (9.3 %)
Pisa	+3.4 (11.0 %)
Arezzo	+2.9 (9.3 %)
Siena	+1.9 (7.2 %)
Grosseto	+5.3 (6.0 %)
Crown	
Provincial Capital (Crown 0)	+5.7 (35.0 %)
Adjacent to the Capital (Crown 1)	+4.8 (21.7 %)
Other municipalities (Crown 2)	+3.1 (43.3 %)
ZALT	
Inland Mountain [IM]	+0.1 (9.6 %)
Coastal Mountain [CM]	+8.2 (4.5 %)
Inland Hills [IH]	+4.5 (53.3 %)
Coastal Hills [CH]	+4.9 (12.7 %)
Plains [P]	+5.6 (19.9 %)
Total Trend	+4.3

5 Detailed Trend of Crown and Altitude Areas

5.1 The Crown Approach

Looking at Figs. 2 and 3, the classification criterion based on crowns seems to be particularly interesting. Figure 2 shows the population trend of the whole Region in absolute values stratified by crown; Fig. 3 shows crowns quotas over time. Both graphs witness a huge population increase in Provincial Capitals from World War II to 80's. Vice-versa municipalities adjacent to capitals performed a three-step increase in absolute values (1861–1881; 1891–1961; 1961–2001), and a constant trend in relative values. Population in other municipalities remains substantially unchanged since inter-war period, even if its quota decreases significantly in the same period (except for a weak gain in the last two censuses).

Various factors could affect this specific pattern of settlement: some of them could be referred to the demographic features of the population, such as different fertility profile or different force of migration by area. Others could be referred to socio-economic features such as the "economic miracle" experienced by Italy during '60s. Unfortunately, lack of detailed data represents a tough hurdle to overcome.

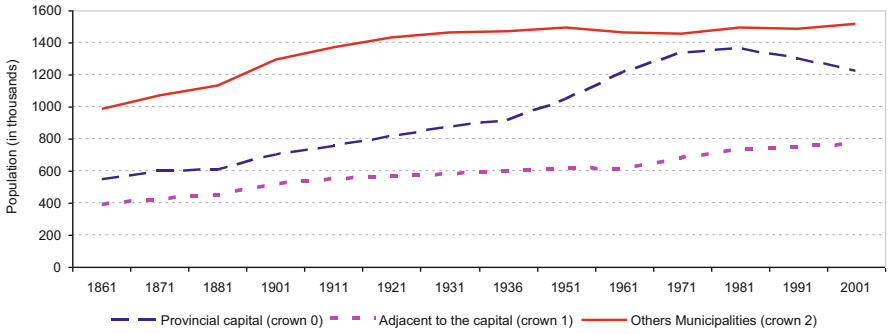

Fig. 2 Population trend by crown, Tuscany (Absolute values)—years 1861–2001

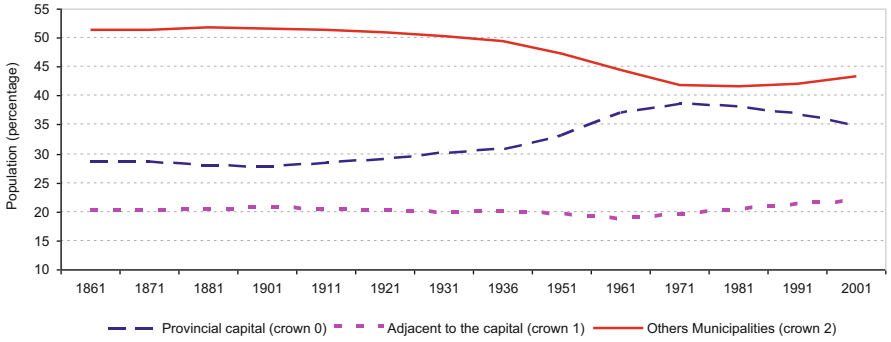

Fig. 3 Population trend by crown, Tuscany (Relative values)—years 1861–2001

5.2 The ZALT Approach

Considering the five altitude/coastal zones, the relative settlement profile of Tuscany population seems to remain quite constant over years.

As showed in Fig. 4, the population quota remained substantially stable in all areas until the inter-war period. After that, the main changes regarded two areas: Inland Mountains and Plains. In particular, Inland Mountains population quota began to decrease constantly over years. On the other hand, Plains areas gained weight until to 20 % of the total amount of population at the last census (2001). This could be an effect of the de-population process experienced by rural areas mostly located in Inland Mountains or on the other point of view, the effect of the urbanization process experienced by urban (and sub urban) areas during the same period.

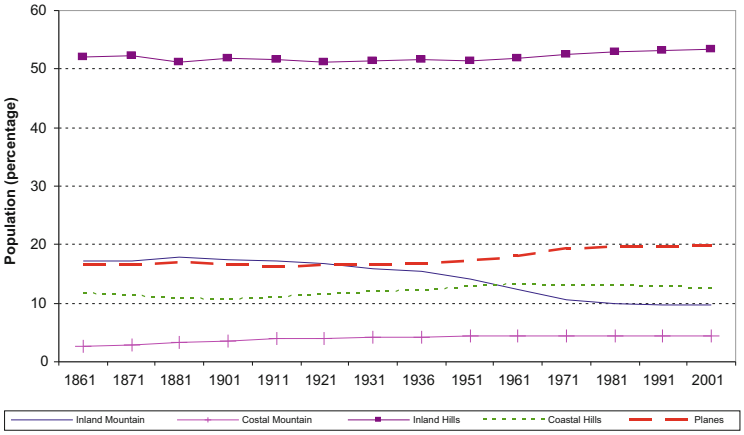

Fig. 4 Total trend of population by altitude areas (*relative values*)

Table 3 Average annual rate of change (per 1,000) between 1861 and 2001 by crown and ZALT

Crown	ZALT				
	Inland mountain	Coastal mountain[a]	Inland hills	Coastal hills	Plains
Provincial capital	3.6	–	6.1	3.2	4.8
Adjacent to the capital	−0.3	–	4.9	5.8	3.8
Other municipalities	−0.8	–	2.6	5.4	6.7

[a]Coastal mountain rates are not included in the analysis due to the small amount of municipalities

5.3 The Crown-by-ZALT Approach

At a deeper level of analysis, a contingency table has been defined cross-tabulating each population census by ZALT and crown profile. After that, the average annual rate of change between 1861 and 2001 has been calculated (Table 3).

The above table shows that population increases over years, except for two cases in the Inland Mountains (Adjacent to capital, other municipalities), promoting the idea to control the specific time series.

While CM, IH, CH, P time series (population profile by crowns) have a similar trend to the whole Region, the ZALT area IM, with reference to other municipalities, shows (Fig. 5) an increasing population trend from 1861 till the inter-war period 1921–1931, and subsequently a general decreasing trend (stronger in the period 1951–1981).

Furthermore, the analysis shows a substantial homogeneity either considering the overtime population change by ZALT or considering both altimetry and crowns in absolute values.

More specifically, following the crown approach it has been possible to highlight how the geographical location is able to affect considerably the settlement profile of the whole population. Introducing the altimetry/coastal dimension, the main features

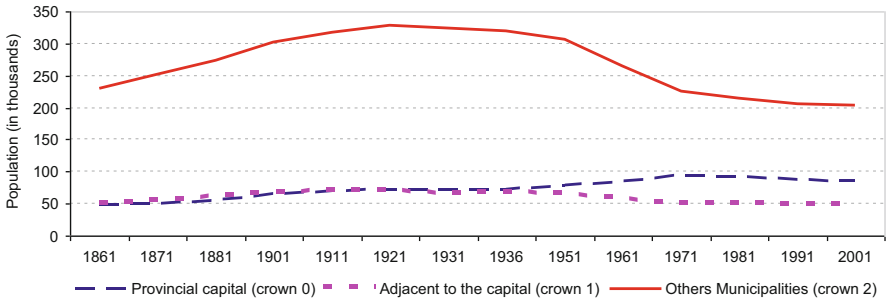

Fig. 5 Total trend of population by crown for Inland Mountain (IM) Municipalities (*figure in thousands*)—years 1861–2001

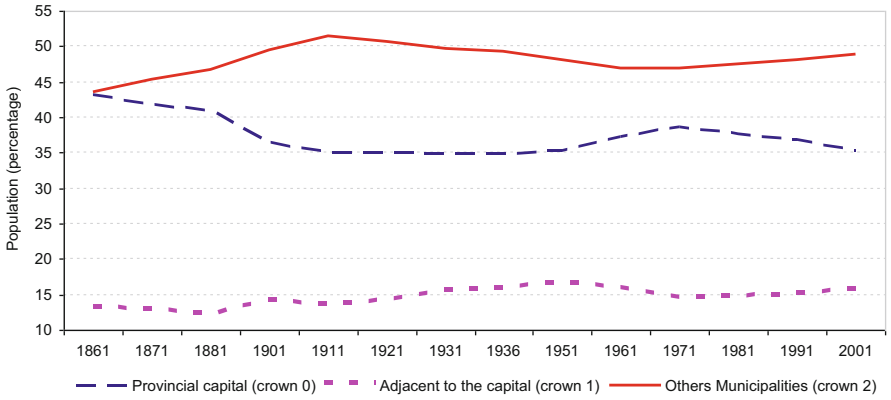

Fig. 6 Distribution of Coastal Hills [CH] by crowns (*figure in percentage*)—years 1861–2001

observed for the whole Region remain stable for the three geographic areas: Inland Hills, Coastal Hills, and Plains.

Shifting from absolute values to relative ones, it is possible to analyze changes in quotas over time. More in detail only two ZALT, Inland Mountains and Inland Hills, are able to explain a trend comparable to the one previously described. Instead, as shown in details below, Coastal Hills and Plains display two specific profiles.

5.3.1 Coastal Hills [CH]

Despite their regional quota remains substantially stable (Fig. 4), municipalities classified as Coastal Hills present a non-monotonic increasing trend in Crown 2 mirrored by a general population reduction in Crown 0, composed by Lucca and Pisa (Fig. 6). Instead, population quota belonging to municipalities Adjacent to the Capital remains quite steady. This trend might be due to the presence of several industrial cities such as Empoli, Pontedera, Fucecchio, Ponsacco.

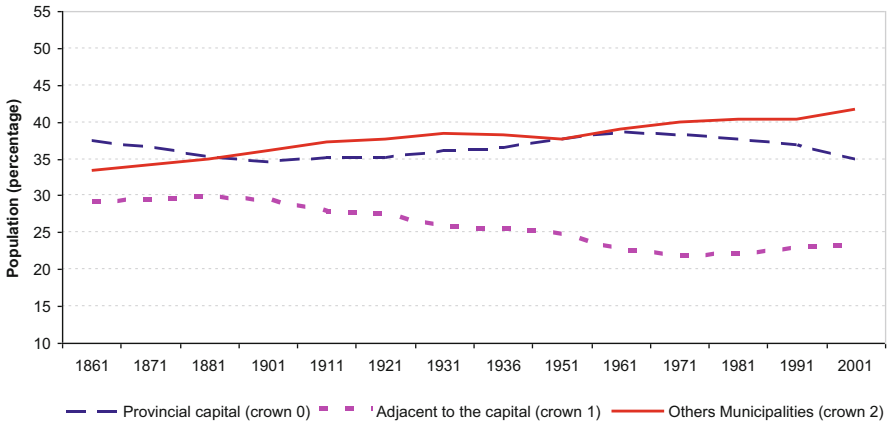

Fig. 7 Distribution of Plains [P] by crown (*figure in percentage*)—years 1861–2001

5.3.2 Plains [P]

As shown in Fig. 4, the regional population quota in municipalities located in Plains increases steadily over years. This growth is mainly affected by crown 2 (Fig. 6) and this is partially balanced by the decrease in municipalities Adjacent to the Capital (crown 1). Livorno, the only Provincial Capital of this geographical division, keeps its relative weight approximately constant (Fig. 7).

6 Conclusions

The study of population trends over years can lead to some interesting results. In particular, classifying data by geographic and administrative variables may show some trends that could be hidden analyzing aggregate data. As showed in Sect. 5, drilling down data provides a deeper level of investigation. Indeed, cross-tabulating the number of inhabitants by crown and ZALT provides a more detailed understanding of the overtime trends of distribution.

This work showed that population settlement profiles in Inland Mountains and Inland Hills are qualitatively similar to the whole Regional profile. Instead, Coastal Hills and Planes are going to experience a continuous population increase, in relative values, in cities classified in other municipalities; this increase is part of a general process of population growth of the whole geographical area.

Clarifying some reasons of this process is not an easy task to cope with, particularly because of lack of data. Higher cost of living in Provincial Capital, difficulties or facilities to get the capital, work constrains, industrialization, economic development, population structure, changes in people preference of settlement could be good reasons able to explain this behavior. Nonetheless, crown approach seems to be able to help gathering geographical differences among areas highlighting heterogeneities.

References

Faustini, L., Sanna, G., Tessitore, C., Valentini, A.: Changes in the geographical distribution of inhabitants in the municipalities of Tuscany during the least 150 years: some empirical evidences. Book of abstract "Statistics in the 150 years from italian unification", Quaderni del Dipartimento di Scienze Statistiche, Serie ricerche 2011, n. 2, Alma mater Studiorum, Università di Bologna (2011)

Istat: Unità amministrative – Variazioni territoriali e di nome dal 1861 al 2000 (Popolazione legale per Comune ai censimenti dal 1861 al 1991) (2001)

Istat: Popolazione residente dei comuni - Censimenti dal 1861 al 1991 (Circoscrizioni territoriali al 20 ottobre 1991) (1994)

Porciani, L., Valentini, A.: A new web information system on demographic census data: the case of Tuscany (SITO.DEM). In: Poster presented at Youpop conference, Firenze (2011)

WHO Regional Office for Europe, European health for all database, Copenhagen, Denmark. http://www.euro.who.int/hfadb

The "Administrative" Territory from the Unity of Italy to the Present

Orietta Gargano and Tiziana Clary

Abstract

This work is intended to bean analysis of the historical changes occurred over time and space in the constitution of municipalities and provinces (administrative units) from the Unification of Italy up to now. For the first time in the last few years a systematic and accurate work was carried out for a full recovery of sources and acts tracing the historical path of the Italian administrative units.

All data were collected and organized in an information system providing their integrated management and availability on-line.

1 The Evolution of the Territory and Administrative Units

At the time of the unification of Italy, the administrative organization model of the Kingdom of Sardinia, established by Rattazzi Decree (Law no. 3702, October 23rd 1859), was extended to the whole Kingdom of Italy, which from the administrative standpoint was divided into Provinces, Districts, Surroundings and Municipalities. At the time of the first census in the history of our Country, that was on December 31st 1861, the national territory corresponded to the current 59 provinces and 15 departments, now regions (see Fig. 1). The municipalities were 7,720 (see Table 1), while the total area, according to the present provincial boundaries, amounted to 256,240 km^2.

This framework has evolved through different historical periods. Following the war between the Kingdom of Italy and the Austro-Hungarian Empire (the third war of Independence), in 1866 the territory of Veneto—Friuli included—and the

O. Gargano (✉) • T. Clary
Istituto nazionale di statistica, Rome, Italy
e-mail: gargano@istat.it; clary@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_9,

Fig. 1 Italy provinces at the date of first census in 1861

Table 1 Number of existing municipalities at the time of censuses classified by geographic areas

Geographic areas	1861	1871	1881	1901	1911	1921	1931	1936	1951	1961	1971	1981	1991	2001	2011
North-west	4,064	3,769	3,686	3,681	3,699	3,702	2,689	2,689	2,960	3,057	3,064	3,064	3,064	3,061	3,059
North-east	364	1,118	1,115	1515	1,125	1,968	1,436	1,429	1,412	1,484	1,482	1,481	1,481	1,480	1,480
Centre	707	927	901	907	915	927	937	943	982	992	998	1,000	1,001	1,003	996
South	1,855	1,840	1,836	1,838	1,860	1,873	1,625	1,648	1,752	1,771	1,638	1,787	1,789	1,790	1,790
Isles	730	728	721	721	724	724	624	630	704	731	738	754	765	767	767
Italy	7,720	8,382	8,259	8,262	8,323	9,194	7,311	7,339	7,810	8,035	8,056	8,086	8,100	8,101	8,092

province of Mantua were annexed. After the union of the former Hapsburgic provinces (Belluno, Padua, Rovigo, Treviso, Venice, Verona, Vicenza) and the annexation of Rome in 1870, the total number of provinces amounted to 69 (Direzione Generale della Statistica 1889).

The provinces of the present Trentino-Alto Adige and Venezia Giulia were annexed in 1920, while three new provinces (La Spezia, Trieste and Ionian) were established in 1923. In 1924, with the annexation of Zara (Zadar), Istria (Pula) and Carnaro (Rijeka)—a total land area of 8,953 km^2 the number of provinces raising to 76, Italy reached its largest territorial extent (Direzione Generale della Statistica 1925).

This extension would permanently settle down with the end of World War II, in 1947. After the Treaty of Paris, a total of 7,625 km^2 were surrendered to foreign state and Italy suffered major territorial losses mainly in favour of ex-Yugoslavia and, to a lesser extent, of France.

In the north-east, Venezia Giulia, Istria and Dalmazia lost 105 municipalities: 33 municipalities were detached from the province of Gorizia; 15 from Trieste (whose territorial district, with the remaining 6 municipalities, responded to the boundaries of Zone A of the Free Trieste territory, under the allied forces administration); Fiume and Zara provinces were suppressed having lost 13 and 2 municipalities, respectively. As a result of the Peace Treaty, even the province of Istria (Pula) was extinguished by ceding 31 municipalities to the ex-Yugoslavia while the remaining 11 municipalities were incorporated within the Zone B boundaries of the Free territory of Trieste, which being under the control of the Yugoslav administration would no longer be subject to census survey. In the north-west it was Piedmont to suffer losses in favour of France. The province of Turin lost 6 municipalities, while Cuneo gave up Briga and Tenda as well as parts of the land belonging to 3 other municipalities (Istituto Centrale di Statistica del Regno d'Italia 1939-XVII).

At the present time, the total land area of Italy is of 301,336 km^2.

Once consolidated the national boundaries, only the internal disposition of the administrative units, relating to the number of provinces and municipalities, has been modified.

In the years of fascism, in particular, the internal outline of the administrative boundaries underwent profound changes, with shifting, merging, suppressing municipalities and establishing new provinces. The most significant event was recorded in 1927 when, following the Royal Decree no. 1, January 2nd 1927, 17 new provinces were established and the province of Caserta was suppressed (though re-established in 1945).

By the same Decree, in 1927, all the districts which had undergone downsizing since 1926 were definitively abolished. In the first decade of the twentieth century, by the R.D. no. 554, May 19th 1912, and R.D. no. 148, February 4th 1913, the districts commissariats still existing in Veneto and Mantua provinces were abolished and transformed into unified districts, while the R.D. no. 1890, October 21st 1926, decreed the elimination of 94 districts and, consequently, of the less important sub-prefectures (Istituto Centrale di Statistica del Regno d'Italia 1927).

At the birth of the Republic, there were 91 effective provinces in Italy.

As regards Aosta province, the power was transferred to the newly formed region with special status, in 1948. In 1974, even considering Aosta Valley, 95 provinces could be counted. Their number increased further on with the establishment of eight new provinces, in 1992; four new provinces became operative in Sardinia in 2005 and finally, three more units (Monza e della Brianza, Fermo, Barletta–Andria–Trani) were added in 2009. At the present time, the provinces are 110 (see Fig. 2).

There were 8,382 municipalities in 1871 (see Table 1). Through various fluctuations they reached 9,194 units in 1921. In the following decade 2,189 municipalities were suppressed (23.8 %), then they increased again in 1951 (+6.8 % compared to 1931) and they amounted to 8,035 units in 1961 (+2.9 % over the previous decade). Since then there has been a slight but steady increase up to a substantial stabilization starting from 1991 onward. Dating from March 1st 2011, the municipalities are 8,092.

1.1 Administrative Variations

Over time, from an administrative standpoint, there has been a rather lively dynamics involving the actual existence of municipalities, the establishment of municipalities and provinces, the setting up of new regions, naming changes—variously concerning all types of administrative entities—and variations in municipality districts (acquisition and/or disposal of land portions).

Any variation occurring and being recorded is bound to a formal act ratifying it and identifying its starting (temporal, or administrative, validity). Six different types of variation have to be pointed out: establishment and/or annexation to the national territory (CS); suppressions (ES); land disposal and acquisitions between municipalities (CE/AQ); variations in the composition of provinces (AP); denomination changes (CD). See Fig. 3.

Over the 150 years of life of our Country since the Unity, the geographical area with the highest number of territorial and administrative changes, occurred and known, has been the north-west (6,776) followed by the north-east (3,320), the Centre (1,632), the South (2,302) and the Islands (1,173).

From a regional standpoint Lombardy shows the highest numbers of variations (3,525), with a record for each type of classified change, exception made for changes in the belonging to province and/or region (AP), where Piedmont stands high (529 events), followed by Venezia Giulia (516). The region with the lowest values is Basilicata, with 120 occurrences.

As refers to provinces, at present the sum of different types of changes shows the predominance of the Autonomous Province of Trento[1] with 701 occurrences, while the province of Prato, with 9 events, has the lowest rate.

The distribution of variations by type, year and territorial reference has not a uniform trend either in time or territory.

[1]In the computation of the variations we have to take into accout that Trento first belonged to Venezia Tridentina compartment, later to Trentino-Alto Adige region and is finally an autonomous province.

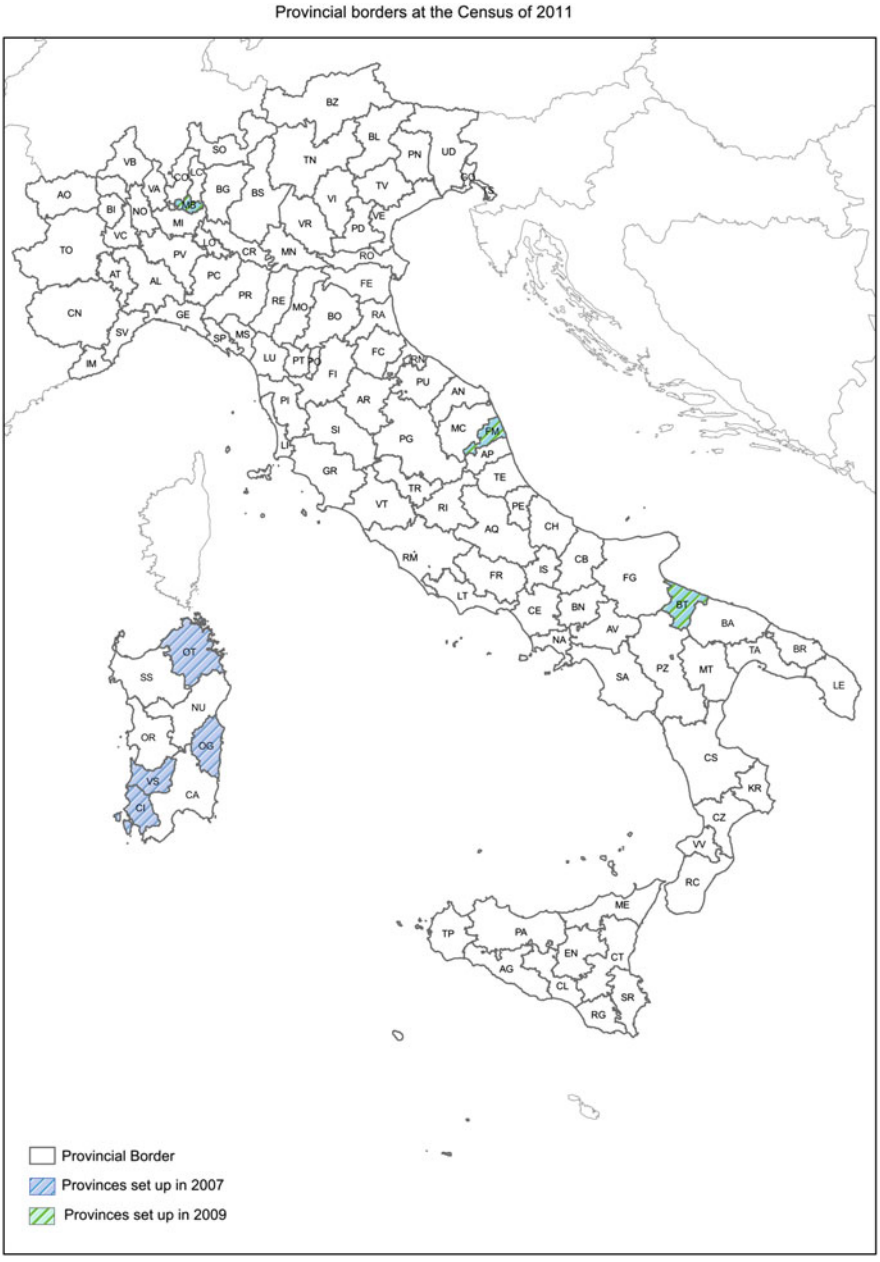

Fig. 2 Italy provinces at the date of the last census in 2011

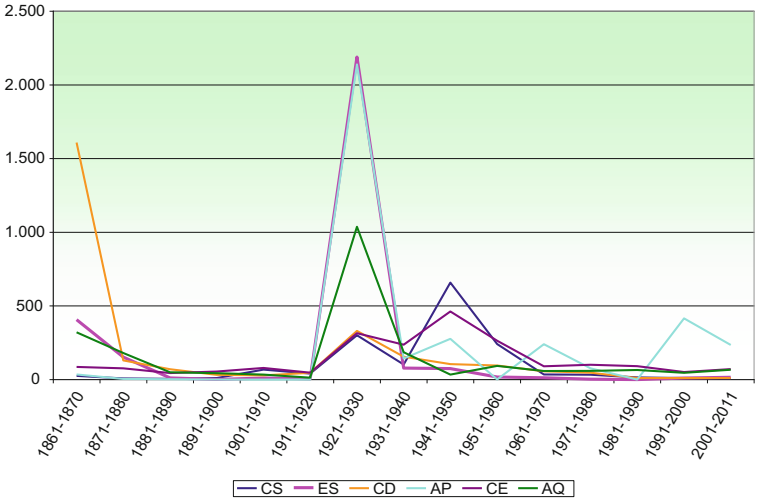

Fig. 3 Variations range from 1861 to 2011

There are two main times standing out with reference to the amount of occurred changes. The first corresponds to the decade between the establishment of the Kingdom of Italy and the proclamation of Rome as capital of the kingdom (from 1861 to 1871). In this period many changes were applied in order to reach the administrative uniformity of the Country.

The second period lasted almost twenty years, coinciding to Fascism government (from 1922 to 1943). In this period lots of laws were inacted in order to reform local government and strengthen the centralizing role of the state (Direzione Generale della Statistica e del lavoro 1911).

These were the years which witnessed: a deep reorganization of the administrative territory, due to the redesigning of Provinces and Districts, in order to strengthen the role of the Prefect; the establishment of the Potestà, a chief executive, appointed by the King, in place of the Mayor.[2]

Denomination changes (see Table 2) took place mostly in the years from 1861 to 1870 (1,609 cases, 59 % of the total). The geographical area with the highest occurrence was the north-west (707 events); the highest figure (394) in Lombardy (Direzione Generale della Statistica 1889, 1900).

Municipalities suppression (see Table 3) was also a significant phenomenon in the above said decade (407 cases, 13,6 % of the total) but the peak was reached in the years from 1921 to 1930: 2,189 cases, for the most part in the north (1,199 in the north-west; 601 in the north-east, of which 384 in Venezia Tridentina, now

[2]The Law no. 237 of 4th February 1926, decreed the Podestà take-over of municipalities with a population of less than 5,000 inhabitants and the R.D.L. no. 1910 of September 3rd 1926 extended the order to all municipalities.

Table 2 Numbers of denomination changes classified by geographic areas

Geographic areas	1861–1870	1871–1880	1881–1890	1891–1900	1901–1910	1911–1920	1921–1930	1931–1940	1941–1950	1951–1960	1961–1970	1971–1980	1981–1990	1991–2000	2001–2011
North-west	707	24	22	14	15	20	88	81	61	40	21	30	1	2	3
North-east	229	12	18	4	5	11	194	34	11	30	18	11	3	2	5
Centre	146	65	12	4	3	8	15	13	14	10	5	2	1	–	2
South	432	20	14	10	6	6	23	17	13	9	9	2	3	3	–
Isles	95	9	4	–	–	1	10	10	6	6	3	2	2	3	–
Italy	1, 609	130	70	32	29	46	330	155	105	95	56	47	10	10	10

Table 3 Number of suppressions of municipalities classified by geographic areas

Geographic areas	1861–1870	1871–1880	1881–1890	1891–1900	1901–1910	1911–1920	1921–1930	1931–1940	1941–1950	1951–1960	1961–1970	1971–1980	1981–1990	1991–2000	2001–2011
North-west	322	108	7	–	4	5	1,199	32	45	14	2	–	–	5	6
North-east	15	6	1	–	–	–	601	19	13	2	7	2	–	4	10
Centre	48	26	2	1	1	–	74	9	3	1	0	1	–	1	–
South	17	6	–	–	2	–	196	10	11	–	3	–	–	–	–
Isles	5	7	1	–	3	–	119	8	3	–	–	–	–	–	–
Italy	407	153	11	1	10	5	2,189	78	75	17	12	3	–	10	16

Trentino-Alto Adige), (Direzione Generale della Statistica 1925 e Istituto Centrale di Statistica del Regno d'Italia 1930).

The decade 1921–1930 even holds a record regarding changes in belonging to provinces or to regions: 2,139 cases, 60 % of the total, a result mostly due to the creation of 17 new provinces, in 1927, involving the whole national territory (see Table 4), (Istituto Centrale di Statistica del Regno d'Italia 1934, 1937).

The second high value (311 municipalities) is to be found in more recent years (1991–2000) when eight new provinces were established, largely in the north-west.

The setting up of new municipalities was more frequent through the years 1941–1950: 658 new municipalities were set up, of which about half in the north-west (333), 103 in the north-east and 103 in the South (see Table 5). In the post-war period, this phenomenon is linked to the cancellation of the fascist measures and the restoration of a democratic order (Istituto Centrale di Statistica 1950).

Such an occurrence can also be observed in the periods 1921–1930 and 1951–1960. In both times, a greater number of new municipalities were in the north-west. Territorial variations or exchange of territories between municipalities are not a prominent phenomenon. Yet, even acquisitions show a very high frequency in the decade 1921–1930 (Istituto nazionale di Statistica 2001).

In that period of time, assignments ranked second in frequency while the largest amount of land disposal is remarked from 1941 to 1950.

In short, the density of events is basically focused in the twenties, particularly in the north; only as regards name changes, the maximum density occurs in the period 1861–1870, particularly in the north-east, with 43.9 %.

Variability, however, showed a downward trend from the end of the war to more recent years. Starting from 1960 there has been some stability both in the distribution of administrative units and in their territorial composition. The only exception is the change in the belongings of municipalities to provinces and regions, a phenomenon in turmoil even in more recent decades, due to the establishment of new provinces.

2 Future Prospects

The availability of information on-line[3] will be of great benefit to researchers and practitioners making it possible for them to know in real time the exact consistency of administrative units and the "n" decision made over time for each of them.

Last step, but not least, is the pursuit of an objective not achieved yet, though planned, that is the unique coding of all the instances taken by the administrative units of the system, with reference to the relevant administrative area at the time of the change. The code, structured, will have a purely statistical significance and will contain references to the year of variation, the Istat coding standard of the territorial administrative unit and the type of variation.

[3]See the following link: http://www.istat.it/it/archivio/6789; http://sistat.istat.it/sistat/.

Table 4 Number of variations in the composition of provinces by geographic areas

Geographic areas	1861–1870	1871–1880	1881–1890	1891–1900	1901–1910	1911–1920	1921–1930	1931–1940	1941–1950	1951–1960	1961–1970	1971–1980	1981–1990	1991–2000	2001–2011
North-west	35	3	–	–	–	–	711	109	65	–	–	–	–	311	55
North-east	–	–	3	–	–	–	636	–	129	–	52	–	–	20	7
Centre	–	–	–	–	–	–	396	31	1	–	–	–	–	7	40
South	–	1	–	–	–	–	206	3	82	–	188	1	–	77	10
Isles	–			–	–	–	133	–	–	1	0	75	–	–	123
Italy	35	4	3	–	–	–	2,139	143	277	1	240	76	–	415	235

Table 5 Number of establishment of municipalities by geographic areas

Geographic areas	1861–1870	1871–1880	1881–1890	1891–1900	1901–1910	1911–1920	1921–1930	1931–1940	1941–1950	1951–1960	1961–1970	1971–1980	1981–1990	1991–2000	2001–2011
North-west	11	7	1	1	21	10	192	15	333	111	9	–	–	2	4
North-east	4	–	–	1	9	4	49	9	103	76	4	2	–	3	3
Centre	4	–	4	5	9	12	13	16	39	10	7	3	1	2	1
South	3	1	–	2	24	11	33	49	103	19	6	13	2	1	–
Isles	3	–	–	1	5	4	14	13	80	25	9	15	12	2	–
Italy	25	8	5	10	68	41	301	102	658	241	35	33	15	10	8

References

Direzione Generale della Statistica – Ministero di Agricoltura, Industria e Commercio: Variazioni nel nome, nel territorio o nella dipendenza amministrativa dei comuni, dei circondari (o distretti) e delle province avvenute dal 1° gennaio 1862 al 31 dicembre 1888. Roma, Tipografia Fratelli Centenari (1889)

Direzione Generale della Statistica – Ministero di Agricoltura, Industria e Commercio: Variazioni avvenute nelle circoscrizioni amministrative del Regno dal 1° gennaio 1882 al 31 dicembre 1899. Roma, Tipografia Nazionale di G. Bertero (1900)

Direzione Generale della Statistica e del Lavoro – Ministero di Agricoltura, Industria e Commercio – Ufficio Centrale di Statistica: Variazioni nelle Circoscrizioni amministrative del Regno avvenute nell'intervallo fra il Censimento del 10 febbraio 1901 e quello del 10 giugno 1911. Roma, Tipografia Nazionale di G. Bertero (1911)

Direzione Generale della Statistica – Ministero dell'Economia Nazionale: Variazioni di territorio e di nome avvenute nelle circoscrizioni amministrative del Regno durante il periodo fra il V e il VI censimento (10 giugno 1911 – 1° dicembre 1921) e il periodo dal 1° dicembre 1921 al 31 dicembre 1924. Roma, Libreria dello Stato (1925)

Istituto Centrale di Statistica del Regno d'Italia: Variazioni di territorio e di nome avvenute nelle circoscrizioni amministrative del Regno dal 1° gennaio 1925 al 31 marzo 1927. Roma, Stabilimento Poligrafico per l'Amministrazione dello Stato (1927)

Istituto Centrale di Statistica del Regno d'Italia: Variazioni di territorio e di nome avvenute nelle circoscrizioni comunali e provinciali del Regno dal 1° aprile 1927 al 15 ottobre 1930. Roma, Tipografia Operaia Romana (1930)

Istituto Nazionale di Statistica: Popolazione e circoscrizioni amministrative dei comuni. Variazioni territoriali e di nome dal 1° gennaio 1950 al 31 dicembre 1954. Roma, Tipografia F. Failli (2001)

Istituto Centrale di Statistica del Regno d'Italia: Variazioni di territorio di nome e di confine delle circoscrizioni comunali e provinciali del Regno disposte con Leggi e Regi decreti emanati dal 16 ottobre 1930-VIII al 31 marzo 1934-XII. Roma, Istituto Poligrafico dello Stato (1934)

Istituto Centrale di Statistica del Regno d'Italia: Variazioni di territorio di nome e di confine delle circoscrizioni comunali e provinciali del Regno disposte con Leggi e Regi decreti emanati dal 1° aprile 1934-XII al 30 aprile 1936-XIV. Roma, Istituto Poligrafico dello Stato (1937)

Istituto Centrale di Statistica del Regno d'Italia: Variazioni delle circoscrizioni comunali, provinciali e delle zone agrarie dal 21 aprile 1936 – XIV al 31 dicembre 1938-XVI. Spoleto, Arti Grafiche Panetto and Petrelli (1939–XVII)

Istituto Centrale di Statistica: Variazioni di territorio e di nome delle circoscrizioni amministrative e delle zone agrarie dal 1° gennaio 1939 al 31 dicembre 1949. Roma, Tipografia Fausto Failli (1950)

Fifty Years of Business Confidence Surveys on Manufacturing Sector

Bianca M. Martelli, Giancarlo Bruno, Paola M. Chiodini, Giancarlo Manzi, and Flavio Verrecchia

Abstract

In this work the evolution of the Italian Business Confidence Survey on manufacturing sector is presented starting from the preliminary European project for harmonized statistics launched in the late fifties of the last century. Survey changes are described, focusing in particular on the so-called *confidence indicator*. The continuing increase of statistical accuracy in sampling is recalled, from the initial purposive sample and controls, up to the present state of the art. Specific attention is devoted to the role of administrative archives in the sampling plan. Emphasis is also given to the increasing use of computer simulation in assessing the validity of the estimates. The role of cyclical analysis is finally highlighted with regard to two aspects: (1) the business confidence has not a corresponding variable in the economic system—the validation can only be performed in comparison with correlated variables (e.g. IP, GDP); (2) confidence shows forecasting capability for the economic system.

B.M. Martelli (✉) • G. Bruno • F. Verrecchia
ISTAT, Rome, Italy
e-mail: bmartelli@istat.it

P.M. Chiodini
Department of Quantitative Methods for Economics and Business, University of Milano-Bicocca, Milan, Italy

G. Manzi
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_10,

1 The Harmonized Business and Consumers Survey: History and Characteristics

The survey on the manufacturing sector in Italy is part of the Joint Harmonised Business and Consumers Survey (BCS) program of the European Commission which presently covers manufacturing, construction, retail trade, services sectors and consumers in all the member countries. About 60 years ago an innovative project was started by the European Commission with the purpose of monitoring the confidence of the economic agents collected in a simple and effective way, i.e. through qualitative opinion surveys performed with monthly frequency. The project gradually involved all the European countries as to currently comprise all the 27 member states.

With this regard, the European Commission states that "the principle of harmonisation underlying the project aims to produce a set of comparable data for all European countries" (EC 2006). To achieve this goal institutes must respect two basic principles: (1) to use the same harmonized questionnaire; (2) to strictly respect the Commission timetable in carrying on the survey and transmitting the results. On the other hand, statistical institutes are left relatively free to define the other aspects of the entire process from data collection to sample design (apart from a required minimum sample size) and processing techniques. They are also invited to conform to the recently developed EC–OECD guidelines (EC 2006; OECD 2003).

The BCS aims to investigate the confidence of the economic operators by asking entrepreneurs and managers on current economic and business trends and expectations for the near future. Information collected is qualitative, mainly on a three-option ordinal scale, whose values (e.g. "above normal", "normal", "below normal"; "high", "normal", "low", etc.) may be sorted into a sequence without any ambiguity. Moreover, possible answers are always presented along with the "I don't know/non-response" option. In some restricted cases, for variables that are not reported in conventional statistics, information collected is quantitative (percentages of capacity utilization; number of months of production assured; etc.).

Answers obtained from the survey are aggregated in the form of *balances* that is as differences between positive and negative answers. Balances are then used to build the *confidence indicator* as arithmetic mean of three series: level of orders, production expectation and stocks (with inverted sign). The general idea behind the construction of such an indicator is that each survey answer contains a common component which can be better extracted by a cross-sectional average. The series, stemming from the monthly information, represent a valuable tool for cyclical analysis and for building leading indicator of the industrial production and the GDP.

In Italy, this survey has a very long history and has always been embedded in the European Project. ISCO[1] (merged in 1999 in ISAE[2] and in 2011 in ISTAT[3]) was among the three statistical institutes (with IFO for Germany and INSEE for France) which started the project in 1959, on a quarterly basis. The survey became monthly based in 1962 on a limited number of questions (ISCO 1961). The project continued over the years according to the European guidelines and progressively upgrading the sampling techniques and the sampling design. Since 1988, the data collection mode gradually shifted from ordinary mail to telephone, assuring more up-to-date results. The data processing received two main revisions, in 1986 and in 2002 (Malgarini et al. 2005), whereas the weighting system was based on internal and external weights at stratum level according to the OECD guidelines (OECD 2003). Following the European Commission recommendations, in May 2010 data were re-classified according to Nace Rev.2 classification.

2 Sampling Design

At the beginning, the survey was intended as a purposive panel of *leading firms*. According to this definition, only enterprises which gave some particular innovative contribution to the growth of industrial sectors were considered (Martelli 1998). The unit selection criteria were therefore mainly discretional with low reliability in the estimates. The original sample size was about 2,600 units stratified by a very detailed economic sectors breakdown (i.e. mainly reflecting the NACE 1 three digits classification). This purposive sample structure has been preserved over about 25 years. Since the eighties of the last century, the increasing use of computational methods led in 1986 to a first thoroughly re-designing of the sample by adopting a proportional allocation, which allowed for an estimation of overall regional outcomes (Pinca 1990). The double need to obtain estimates both with sectorial and regional breakdown was dictated by the European project guidelines to collect both country and sectorial data, and by the increasing domestic demand for local information. Both these needs, however, were conflicting with the precision of the domain estimates as the sample size could not be increased due to budget constraints.

As an alternative solution, at least to improve the quality of the overall estimates, further sample designs were tested. In 1998, a univariate x-optimal allocation (Martelli 1998) was applied to a stratified sampling design with 22 macro sectors (according to Neyman-based workforce variance, estimated from previous waves of the survey), 3 firm sizes and 19 Nuts areas (i.e. mainly Nuts-2). This allocation

[1] Institute for Short Term Analysis.

[2] Institute for Economic Analyses.

[3] Italian National Institute of Statistics.

allowed for the calculation of a sampling error of only about ±0.5 % according to the average of the three qualitative questions composing the confidence indicator.

From 1999 onwards, the availability of the business frame ASIA[4] (Statistical Archive of Active Firms) provided by ISTAT (Eurostat 2006; ISTAT 2010) resulted in a significant improvement of several aspects of the survey design, namely: (1) in defining the frame, (2) in unit selection, (3) in variance calculation for the Neyman *x*-optimal allocation,[5] (4) in the sectorial classification, (5) in simulation exercises for testing and validating the sampling design (Chiodini et al. 2010a, b, 2011a, b).

According to (1) above, by using the business frame Asia, under and over coverage problems are now almost completely solved. However, a remarkable time lag persists: ASIA is disseminated about one year and six months later with respect to the information collected.

According to (2) above, the nearly complete information offered by ASIA is an optimal pre-condition for selecting units for the original sample, which usually relies on administrative settings (classifications of economic activities, areas, etc.) and it is likely affected by between-strata heterogeneity (in terms of population size and stratum variance).

According to (3), above, the availability of the business frame ASIA allows for the application of the Neyman allocation to strata using the real variances (on workforce), and not estimated variances drawn for the survey itself (as it was customary in previous attempts).

According to (4) above, on March 2009 the European Commission set the deadline to have all the BCS classified according to the Nace Rev.2 classification. This requirement implied, among others, the revision of the domains (strata) of the survey (Eurostat 2006). To this purpose, the ASIA archive played a determinant role by offering in 2007 the double classification of the firms according both to the old Nace rev.1 (Ateco 2002) and to Nace Rev.2 (Ateco 2007) allowing both for the careful reconstruction of the time series of the results and the revision of the strata.

According to (5) above, only recently researchers have dealt with computational methods and simulations in the field of sample allocation (Chiodini et al. 2010b) as it represents a powerful tool for testing the sample allocation efficiency at a stratum level. This occurrence is useful in the allocation exercise when a high number of strata are required. Furthermore, simulation has an additional important

[4]The ASIA archive is set up and yearly updated by the Italian National Institute of Statistics by merging some main administrative archives that is those of the Italian Economy Ministry, Italian Chamber of Commerce, Italian Social Security (INPS), Italian National Insurance Institute for Industrial Accidents (INAIL), Italian Telephone Company (Telecom) and Italian National Electricity Board (ENEL). ASIA represents the most complete and updated source of the Italian firms' universe. It allows reliable and complete information for both building the sample and selecting the addresses, overcoming the usual problem to have a partial frame list in comparison to the universe. It is disseminated with about one-and-a-half year delay with respect to the information collected. This fact further allows keeping updated information on the universe between the Census Surveys, which are usually carried out every 10 years.

[5]The 1998 sample allocation benefited from the 1996 pilot release of the ASIA archive.

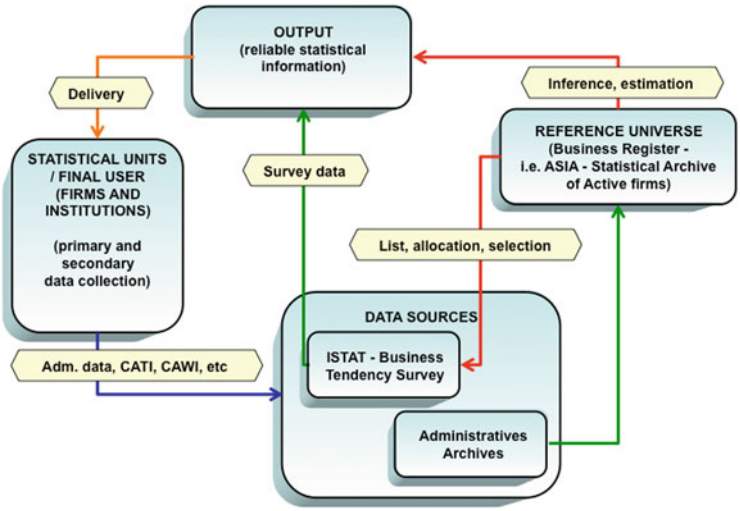

Fig. 1 Current confidence survey design process: actors and actions *Source*: Slide presented at Enhancement and Social Responsibility of Official Statistics, 1st SISvsp Workshop, Rome, April 27–28. See Chiodini et al. (2011b)

feature: as confidence surveys do not have a benchmark in the universe to validate the outcomes, the only possible strategy to evaluate the power of the estimates is offered by simulation tools. It must be noted that in recent literature on this topic there are plenty of proposals for new estimators which are related to the introduction of new methods of sampling unit allocation within population strata, representing a valuable alternative to Neyman's optimal allocation method (see, e.g. Étoré and Jourdain 2010; Kaur et al. 1997), and whose statistical features are validated through intensive Monte Carlo simulation.

In Fig. 1 the Confidence survey is synthetically presented by showing all the components of the entire process and their reciprocal relationships.

The availability of the ASIA archive allowed for the setting of new computer-driven strategies for simulation (when methods and estimate performances have to be simultaneously compared). For example, Chiodini et al. (2010b, 2011a) used a method called *Sequential Selection-Allocation*, which is a sequential process to empirically evaluate the performance of the various sampling allocation methods by constructing a new labeled list with population units re-labeled within the stratum according to their selection order, after performing a Sampling Without Replacement (SWOR) of size equal to the stratum size. This process is repeated n times. From this new labeled population, all the allocation algorithms can be performed and their efficiency evaluated at the same time. In fact, when the availability of real data is scarce (and this is the case when comparing different scenarios) only computational power can support the empirical evidence. In a recent work, Chiodini et al. (2010a) compared several allocation methods for

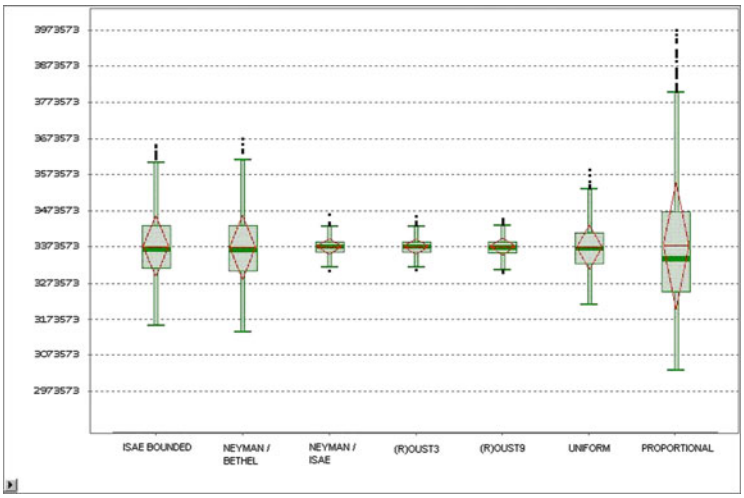

Fig. 2 Total error of the distribution replicates. *Source*: Chiodini et al. (2010a)

the BTS survey (such as the Neyman allocation—currently applied on areas, the Bethel multivariate allocation—as widely applied by ISTAT, now available as a "generalized software", the uniform and the proportional allocations, and a novel method, namely the Robust Optimal Allocation with Uniform Threshold method—ROAUST9, which is a Neyman domain method) by applying the SSA simulation technique, in order to re-think the allocation method to be used in a near future.

Chiodini et al. (2010a) use the statistics on the overall workforce in order to compare the allocation methods, as in their simulation the workforce can be considered a proxy of the data to be collected (investigated). Useful criteria are the Absolute Total Error (|TE|) and the Relative Absolute Total Error (|RTE|), given by:

$$|\mathrm{TE}| = |\mathrm{Bias}| + \sigma_{\mathrm{r}}$$
$$|\mathrm{RTE}| : \text{Relative } |\mathrm{TE}| = \left|\mathrm{Bias}/\mu_{\mathrm{r}}\right| + \sigma_{\mathrm{r}}/\mu_{\mathrm{r}} = \left|\mathrm{Bias}/\mu_{\mathrm{r}}\right| + \mathrm{CV}_{\mathrm{r}},$$

where Bias is equal to $\mu - \mu_{\mathrm{r}}$ (μ is the population mean and μ_{r} the replication mean) and σ_{r} is the standard error (SE) of the replicates.

Both Bias[6]—that refers to systematic errors—and SE—that refers to the precision of the estimators—are lower in the Neyman allocation when applied to the overall population (Fig. 2). While the distribution of replications of all the

[6]It must be noted that in this work our main focus is not on asymptotic properties of the allocation methods. Therefore, given a finite number of replications, high bias levels will denote the unsuitability of the methods conditioning on the choice of the stratification variables and the unit selection mode.

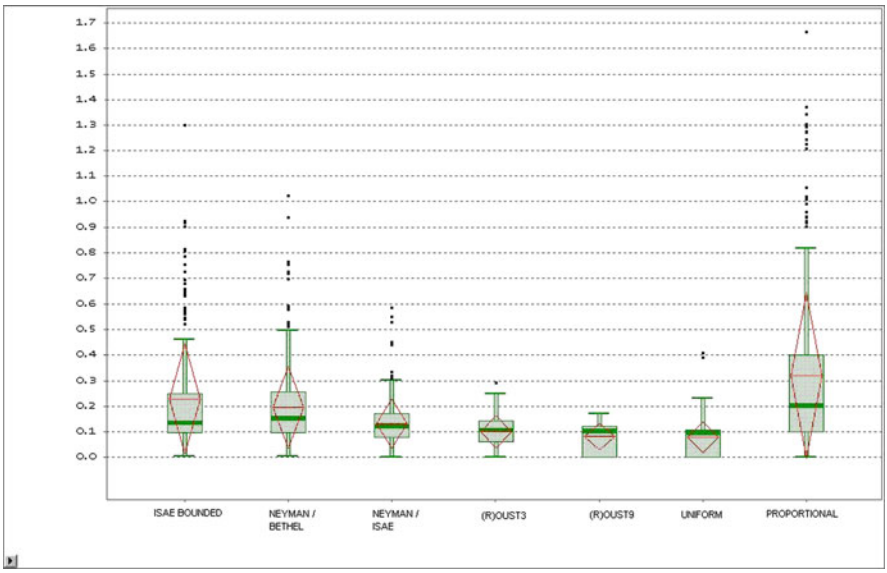

Fig. 3 Relative total error by stratum. *Source*: Chiodini et al. (2010a)

methods based on Neyman's method appears to be centred on the frame mean (i.e. unbiased), the uniform allocation and, at larger extent, the proportional allocation result skewed. Furthermore these two latter methods show a remarkable higher volatility.[7] On the other hand, the ROAUST9 method (although with a little loss in terms of Bias and precision) results to have the higher accuracy within the strata (Chiodini et al. 2010a) (see Fig. 3).

Looking backwards to the first years in which the survey has been carried out, if it is possible from a statistical point of view to accept the purposive sampling selection then performed as a quantitative comparison of quality indicators is out of our reach. A possible validation can in this case arise only ex-post from a cyclical analysis, as it will be shown in the next section.

3 Cyclical Analysis as a Validation Tool

The results from the business survey data need to be validated in order to assess their usefulness as well as their relation with some quantitative indicators. In particular, in this case the industrial production index is a natural candidate for such a comparison. Here we simply consider the comparison between the industrial production index

[7]Better bias and precision levels for the uniform allocation compared to the proportional allocation are connected to an inversely proportional relation between the number of the units within the strata and variability (which is typical of the sectorial and size stratification in business surveys).

and the confidence indicator, even though a more detailed analysis could in principle be carried out also considering the single variables composing the confidence.

A direct comparison of confidence and industrial production, however, would not lead to any meaningful result. In fact, we have to consider that these indicators, while both referring broadly to the activity in the manufacturing sector, nevertheless feature also some subtle differences which must be taken into account while building a possible relation. With respect to this point it is useful to consider the industrial production index as a sum of components: a long term trend, which can be represented by a low order time polynomial; a seasonal, that is the regular movements with period up to a year; the cycle, a recurrent oscillation along the trend with a variable amplitude and periodicity between, approximately, 2 and 10 years; the irregular, i.e. a very short term source of variation not falling in the previous cases.

Considering the confidence indicator, its composing variables are a kind of *diffusion indexes*, defined as the excess of the percentage of firms declaring to face "above the trend" production or order books minus those facing a "below the trend" value (the reverse applies to stocks of finished goods). Therefore, the confidence indicator can also be seen as a diffusion index, capturing what can be thought of as a common component in manufacturing firms' production. This common component is not related to seasonality or long term trends, which are excluded by the definition of the survey question; it is rather likely to represent the cyclical component.

Therefore, the relation between the confidence indicator and the industrial production index will be analysed on the ground of the cyclical behaviour of both series. In order to accomplish this task we will consider various transformations of the industrial production index. A required preliminary step consists in removing its strong seasonal variation, obtaining the so called "seasonal adjusted" series, which here is obtained by means of an unobserved component model (Harvey, 1990).

Indeed, the question we are trying to investigate is whether the business cycle features of the confidence indicator are more related to the concept of *classical*, *deviation or growth cycle* of the quantitative indicator. While the first is consistent with the original definition of business cycle given in Burns and Mitchell (1946) defining a recession as a decline in the *absolute* level of a series, the second and the third are more in line with Mintz (1969) and define a recession, as a decline in the *de-trended* series or, respectively, in the *growth rate* series.

In all the cases the routine proposed by Bry and Boschan (1971) is used to identify the turning points and, therefore, expansion and recession phases. When the classical cycle is considered, business cycle phases are identified directly on the seasonally adjusted industrial production index. In the case of the deviation cycle, it is necessary to specify a suitable de-trending procedure. Due to the fact that turning points detection is highly sensitive to the de-trending method used (Canova 1999) here we rely on two different methods, using the cycles extracted, respectively, with a Butterworth filter (Pollock 2000) and the Hodrick–Prescott filter (Hodrick et al. 1997). These are both low-pass filters for trend estimation, in a series composed by a trend and a cycle component. The filter estimates the trend, while the residual, which is therefore taken to represent the cycle, is considered in the subsequent

Table 1 Correlation between business cycle phases with respect to that of the confidence indicator

	Level	Butterworth	Hodrick–Prescott	Seasonal △ of logs
Correlation at 0	0.210	0.338	0.286	0.487
Max correlation (lag)	0.353 (8)	0.417 (5)	0.321 (2)	0.487 (0)
		Classical cycle	Deviation cycle	Growth rate cycle

Source: Estimations on ISTAT and ISCO–ISAE data

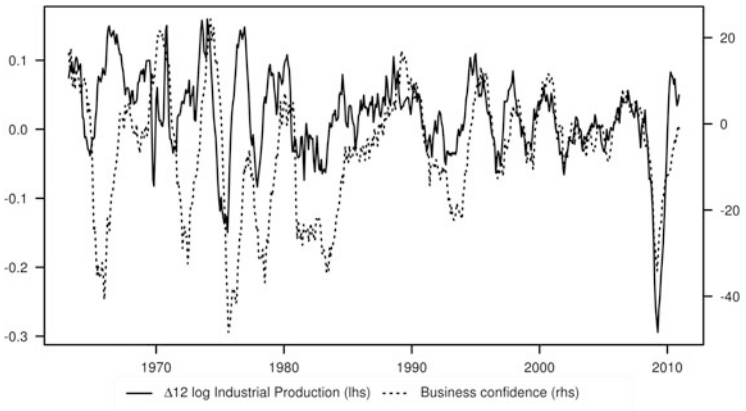

Fig. 4 Confidence and business cycle. *Source*: Estimations on ISTAT and ISCO–ISAE data

analysis. Finally, the growth cycle series considered is the seasonal difference of logs of industrial production.

Once the turning point detection procedure is applied, business cycle phases are represented as binary series, with 1s' representing an expansion and 0s' representing a recession. The relation between the business cycle of the confidence indicator and those of the various transformations of industrial production index are examined with the correlation coefficient, also considering some lagged relationships.

Table 1 reports the main results: the correlation coefficient is reported both for the contemporaneous case as well as for the lead/lag presenting the maximum value. The main facts can be summarized as follows: (1) correlation increases, passing from the classical cycle to the growth cycle, with the deviation cycle somewhat in the middle; this result therefore supports the usual procedure of practitioners of building a relation between seasonal difference of logs of industrial production and confidence indicator for forecasting purposes, given the earlier availability of the latter; (2) in general, there is a lead of business phases for the confidence indicator over the classical cycle and, on a lesser extent, over the deviation one.

The results clearly point out that the concept of growth rate cycle of industrial production is closer to that implied by the confidence indicator.

Confidence (Fig. 4) faithfully tracks the evolution of the Italian economy business cycle turning points, as recorded by the industrial production index, during the

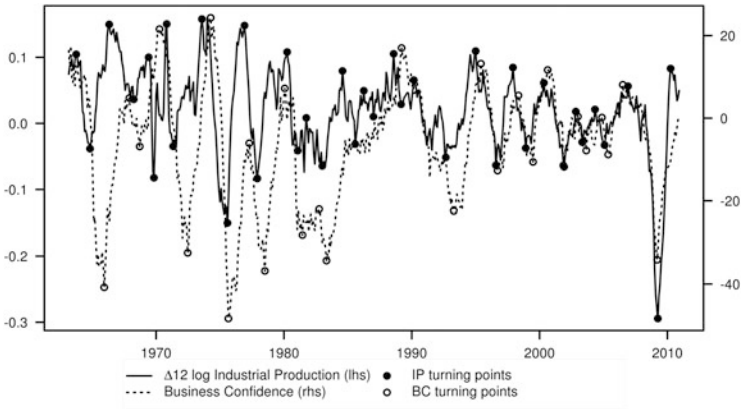

Fig. 5 Estimated turning points. *Source*: Estimations on ISTAT and ISCO–ISAE data

whole period considered, even though the amplitude of business cycle phases does not appear to be always consistent among the two indicators. In the first two decades the shifts of the IP are less precisely recorded by the confidence indicator potentially suggesting the rougher nature of the first sample designs. Starting from the nineties, however, a more marked similarity between the profiles of the two series appears evident. Estimated turning points from the two series are shown in Fig. 5.

4 Concluding Remarks

In this paper we presented the Business Confidence Survey for the Italian Manufacturing sector that was conducted since the sixties of the last century. We synthetically discussed the statistical features of the survey and the improvements occurred over the years. The Confidence indicator is then described and compared to different kinds of economic cycle as recorded by the industrial production index. The paper shows that Confidence faithfully tracks the economic business cycle mainly since the nineties.

From a statistical point of view these occurrences could also support the hypothesis of the effectiveness of the improved sample allocation applied since the nineties (ISAE-Neyman) and give support for the future to the selection of the ROAUST one as suggested by the simulation exercise.

References

Bry, G., Boschan, C.: Cyclical Analysis of Time Series: Selected Procedures and Computer Programs. NBER, New York (1971)

Burns, A.F., Mitchell, W.C.: Measuring Business Cycles. NBER, New York (1946)

Canova, F.: Does de-trending matter for the determination of the reference cycle and the selection of turning points? Econ. J. **109**(452), 126–150 (1999)

Chiodini, P.M., Manzi, G., Martelli, B.M., Verrecchia, F.: The ISAE manufacturing survey sample: validating the Nace Rev.2 sectorial allocation. Paper presented at 30th CIRET conference, New York, 13–16 October 2010a. https://www.ciret.org/conferences/newyork_2010/papers/upload/p_45-185305.pdf

Chiodini, P.M., Lima, R., Manzi, G., Martelli, B.M., Verrecchia, F.: Criticalities in applying the Neyman's optimality in business surveys: a comparison of selected allocation methods. In: Wywiał, J., Gamrot, W. (eds.) Survey Sampling Methods in Economic and Social Research, pp. 42–77. Katowice University of Economics Publishing Office, Poland (2010b)

Chiodini, P.M., Manzi, G., Martelli, B.M., Verrecchia, F.: On computational aspects of simulation methods in the sample allocation framework. Paper presented at 4th ESRA conference, Lausanne, 19–22 July 2011a

Chiodini, P.M., Manzi, G., Martelli, B.M., Verrecchia, F.: Archive and sampling information: is an integration possible? Paper presented at enhancement and social responsibility of official statistics, 1th SISvsp workshop, Rome, 27–28 April 2011b

Étoré, P., Jourdain, B.: Adaptive optimal allocation in stratified sampling methods. Methodol. Comput. Appl. Probab. **12**, 335–360 (2010)

European Commission: The Joint Harmonised EU Programme of Business and Consumer Surveys. In: European Economy, Special Report No. 5, Bruxelles (2006)

Eurostat (Taskforce on the implementation of NACE Rev.2): handbook on methodological aspects related to sampling designs and weights estimations, Version 1.0. (2006)

Harvey, A.: Forecasting, Structural Time Series and the Kalman Filter. Cambridge University Press, Cambridge (1990)

Hodrick, R.J., Prescott, E.C.: Postwar U.S. business cycles: an empirical investigation. J. Money Credit Banking **29**(1), 1–16 (1997)

ISCO, Progetto per un'inchiesta congiunturale Rapida Mensile tra i sei Paesi della Comunità Economica Europea, Congiuntura Italiana, n. 12, Istituto Nazionale per lo Studio della Congiuntura, Rome (1961)

ISTAT: Classificazione delle attivitá economiche Ateco 2002 derivata dalla Nace Rev. 1.1. , Anno 2003, Avellino (2004)

ISTAT: Classificazione delle attivitá economiche Ateco 2007 derivata dalla Nace Rev. 2, Anno 2009, Rome (2009)

ISTAT: Struttura e dimensione delle imprese Archivio Statistico delle Imprese Attive (Asia), Anno 2008, Rome (2010)

Kaur, A., Patil, G.P., Taillie, C.: Unequal allocation models for ranked set sampling with skew distributions. Biometrics **53**, 123–130 (1997)

Malgarini, M., Margani, P., Martelli, B.M.: New design of the ISAE manufacturing survey. In: JBCMA, vol. 2, No. 1, pp. 125–142. OECD, Paris (2005)

Martelli, B.M.: Le inchieste congiunturali dell'ISCO: aspetti metodologici. In: Le inchieste dell'ISCO come strumento di analisi della congiuntura economica, Rassegna di lavori dell'ISCO, Anno XV, n. 3, chap. 1, pp. 13–67 (1998)

Mintz, I.: Dating Postwar Business Cycles: Methods and Their Application to Western Germany, 1950–1967. Occasional Paper no. 107, NBER, New York (1969)

OECD: Business Tendency Surveys: A Handbook. OECD, Paris (2003)

Pinca F.: La Regionalizzazione delle Indagini Congiunturali. In: Strassoldo, M. (ed.), L'Analisi della Congiuntura Economica Locale: Modelli, Metodi e Basi informative. CEDAM, Padova (1990)

Pollock, D.S.G.: Trend estimation and de-trending via rational square-wave filters. J. Econ. **99**, 317–334 (2000)

Part III

New Developments in Survey Methodology for Official Statistics

Social Aspects on Censuses and Official Surveys in Italy

Enrica Aureli and Mariangela Verrascina

Abstract

In opening this overview of the statistics of the last 150 years, aiming at researching the social aspects of census surveys and Italian official research, we can represent this set as a stream born in a small scale, extracted in the early years from administrative tasks, and has been enriched through successive inflows becoming a river of a certain size when ISTAT was established, and then a greater river through the help of international institutions. Today we can finally represent the production of social statistics such as the delta of a river (SISTAN) in which each branch (Institutes, administration . . .) contribute to the sediment at the mouth mixing with water (data, information) of others. Three great periods linked to the official statistic can be characterized: the period of the Central Bureau of Statistics (from 1861 to 1926); the period of the Central Statistical Institute (from 1926 to the early 1980s); the period since the early 1980s to the present. Within these periods, it is possible to identify the sub-periods related to events, even outside the official statistics system, but significant in addressing the issues and policies.

E. Aureli (✉)
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
e-mail: enrica.aureli@uniroma1.it

M. Verrascina
Istituto Nazionale di Statistica, Rome, Italy
e-mail: verrasci@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_11,

1 The Period of the Central Bureau of Statistics (From 1861 to 1926)

The first population census of 1861 in Italy responded to the need to provide politicians with the background elements of the new political reality that was created with the birth of the Kingdom of Italy. The aggregation of many different realities was leading to the birth of a nation and to the characteristics of the people who would go on to create it and highlight the specific needs and the potential success of policy measures and administrative systems that would result in the dismantling of the previous administration. Thus, in the same year as the unification of Italy, 31 December 1861, the first census of the population was carried out, with the aim of "defining the basis for the realization of the Kingdom and of perceiving the people's sense of belonging to it." The urgency to know not only the number, but also the natural conditions and civil rights of the people led to critical information in determining many rights and many civic duties, but also the laws governing the most precious of political rights, the electorate. The first census collected personal data such as marital status, age, place of birth from which the electorate could be deduced, and working conditions as well as temporary migration for work. But it also highlighted attention to social issues: religion and spoken languages (in order to identify minorities located in the territory), the level of literacy (in terms of being able to read or write), and the conditions of handicaps (limited to the deaf and blind) (Istat 1957).

A statistical yearbook, edited by Correnti and Maestri, had already been published in Turin by 1858, before the first Census. Such a yearbook presented "the image of Italy as it was in those days, both a servant and divided, but already with an awareness of its unity and full will for a rebirth." Despite the good intentions of the authors, who set out to achieve an annual publication, we had to wait until 1864 for the Second Italian Statistical Yearbook. This yearbook tended to read data and highlight regional differences. Two chapters titled "Medical Statistics" and "Intellectual Italy" were devoted to social issues. The first bore the subtitle "Hygiene of the army," and military doctors were the first to realize the regional differences in the spread of disease. The data on intellectual Italy concerned university teaching, secondary teaching, pupils at primary schools, and a first percentage index calculated relating students to the overall population.

The next census of 1871, 1881, 1901, and 1911 followed the approach of the first one with small changes, adding paternity to the demographic data, omitting the spoken language and even religion in 1881, and implementing, between 1871 and 1901, the disabilities detected. A more accurate specification of the work of the householder required the disclosure of the head of the family's occupation, which drew most of the means of subsistence, and then those of lesser importance is required both in the census of 1881 and in 1911 (Istat 1959). Such specifications foreshadowed the interest in the employment structure in terms of location and economic class and in relation to the possibility of self and family support from segments of income of different importance. Starting from the 1881 census, an interest with respect to foreigners in Italy is shown, and the nationality and the duration of stay in Italy (since 1911) are collected.

During these years the perspective from which we look at the statistics also changes. The "1890 Yearbook of Statistics," published by Brunialti, helps to understand the cognitive interests developed in the meantime in the absence of the 1891 census. The Yearbook is largely set out according to the administrative reorganization of the Kingdom and the events that happened in that year, although the author stated that "1890 will be remembered as a very average year." However, the statistical information section is extended to the Italian press, education, charity, and justice, to which each a chapter is devoted. An appendix to statistics of religions appears, although it refers to the distribution of religions in the world and not to Italy. A brief statement of six pages also appears, describing the economic and social progress of Italy from 1861 to 1889. It summarizes the trends within the above mentioned social fields that are all expressed in absolute values. Interesting to note is the care with which they looked at the press. The growth in the printed media is matched by the increase in literacy and schooling. The result of the enrolled members of the age group of reference highlights one's attention to the use of more appropriate indicators than in the past, so the difference in schooling at the regional level is put in evidence. The interest in public charities highlighted in the yearbook derives from the approval by the Parliament of the relative Law in the same year. However, statistically it could not provide any information yet. The Yearbook provides, however, the condition of the charities present on the territory and the availability in the budget of municipalities and provinces to offer, in the future, a comparative tool with the effects of the new law reforming the sector. Thus one can see a beginning of the use of statistics to assess the success of public policies. The civil and criminal justice data contained are understood as evidence of the moral movement of the country with great detail on the types of cases and proceedings foreshadowing what will from then on be the judicial statistics, all guided by the procedural aspects of administrative provenance. The statistics for criminal justice give an account of the more limited articulation of the offences at the time. Brunialti concludes: "The condition of the city worker has certainly become better. Rents are more expensive but the houses are far better. The price of clothes has diminished, but not that of meat, but that of other foods has decreased. In short, whatever one may say, the economic conditions of workers have come to improve a bit more than the moral conditions." The conclusion would seem a first attempt at analyzing the quality of life.

The 1921 census did not record any significant change in the survey pattern and the information of a social nature that derived from it, as one might have expected as a result of the social changes associated with World War I just finished.

The use of administrative data from the education system allowed for the editing of a volume of "summary of data on the middle and normal school institutes from 1909–10 to 1911–12," whose attention to gender is interesting as an entire chapter is devoted to women present in those institutions.

2 From the Establishment of the Central Institute of Statistics to the Small 1936 Census

A shock to the organization of statistical production in Italy derived from the 16th session of the International Institute of Statistics held in Rome in 1925. Until that time statistical production had been limited to the information needed by individual ministries and government departments. However at that time, the scientific community stepped forward with a strong need for coordination of statistical production that led to the establishment of the Central Statistical Institute, reporting to the Head of Government and replacing the Central Bureau of Statistics that was anchored to the Directorate General of Labour Statistics and the Ministry of Agriculture, Industry, and Commerce. This new collocation that emphasized its role as an instrument for decision making (*numerus rei publicae fundamentum*) exceeds the purely advisory role of the former Superior Council of Statistics.

However, the Census of 1931 and the small census of 1936 follow the establishment of the previous censuses, but show some signs of the kind of information the political power required with more attention. In fact, in 1931 questions were introduced on the age of women at the time of marriage, on any second marriages, on the number of children born, and on the number of children living. This was clearly related to the population policy of the regime. From the perspective of work, the field of economic activity was introduced, and it would be kept in all subsequent censuses. In both censuses, a strong focus was placed on the flows of migration to the colonies or to foreign countries. Such information was of great interest to politicians who were aiming at the colonization of the countries of the empire and strengthening the Italian population numbers seen as a representation of power.

Throughout the period the propensity to systematize the availability of data in education is consolidating. A monograph devoted to the statistics of some Italian cultural events in the period 1931–1935 contains information about libraries, book production, archives, intellectual property, museums and institutions of art, film, radio, and freelance professions. Data reading started using the advanced methods of statistics, and the frequency data is replaced by historical or indices of composition data and data derived from (i.e. pupils studied in relation to sex and type of institutes per compartment). In 1936 a monograph dedicated to students enrolled in universities and high schools in the academic year 1931–1932 was also published perpetuating and expanding on a similar survey conducted in the academic year 1926–1927. The survey presented was very innovative introducing two new dimensions of the study: the spatial mobility of university students according to the attractiveness of the premises and the social class according to their father's occupation and their willingness to follow their father's education or career. Official statistics illustrate the attempts to search for an autonomous role, but serves primarily the function of producing data for the management of public affairs and government policies. The confirmation is in a paragraph dedicated to the students belonging to the "Gioventù Italiana del Littorio" within the statistical data on middle school education for the school year 1936–1937 which included the statistics data from 1932–1933 to 1935–1936.

3 From the Post-War Reconstruction to the Financial Boom: Sample Surveys, Special Surveys

The war resulted in the failure to conduct the census of 1941 and thus led to, in September 1944, the realization of the "Census and Surveys for national reconstruction," carried out in 38 provinces of liberated Italy (in accordance with the commitments made between the Allied Commission and the Presidency of the Council of Ministers). In 2 months were conducted four general censuses (population , agriculture, industry, and merchant marine) and 30 surveys on all major aspects of national life (Istat 1945). Within the Social environment surveys were carried out on living conditions, food, clothing, housing, public health, primary and secondary education, and public services. Implementation difficulties imposed the use of estimates, and the whole experience was a turning point in the approach to produce official statistics as expressed by the ISTAT General Director Molinari, "The usefulness of official statistics is now greater, the more pronounced is its nature of 'topicality' in comparison to those predominantly 'historical': especially when the preparation of 'plans'—which must forcibly be rooted in statistics-occupies a prominent place in the activities of the state."

The first real post-war census of 1951 is confronted with the need to determine what the demographic, economic and social base on which the country's recovery had to be based. The new national territory, involved new borders and a flow of refugees from territories no longer administered by Italy, and the return of refugees from former colonies in Africa, demanded that the survey should include even a question on "Refugees." There were new types of people whose needs and demands the political power had to provide answers and assistance to. In addition, questions on housing and services available that detect the persistence of poor housing, both in relation to crowded conditions and cohabitation and the lack of services were introduced (Cortese 1978).

The Official statistic in this period was aware of its own potentialities and requires going beyond the mere certification of the state of the population. This was also described at intervals which allowed the detection of changes in the long-term without giving an immediate response to any short-term phenomena. In 1952, according to the methodological developments of the discipline and techniques of sample collection, the first sample survey on the Labour Force took place, followed by three more completed (1954, 1956, 1957) and reaching, in 1959, a quarterly survey for a project that would become routine. Such a methodological boost is significantly attributable to the role played by the president of ISTAT Maroi, who designed and created also the dissemination of statistical data according to a new set of systematically issued monothematic directories, mostly on the social field: Yearbook of judiciary Statistics (1949), Yearbook of Italian Education Statistics (1950), Directory Statistics Assistance and Social Security (1951), Yearbook of Health Statistics (1955), Yearbook of Statistics of Emigration (1955, become Yearbook of Labour Statistics and Emigration in 1960). All these could benefit from systematic and comprehensive information from public administrations which helped to strengthen the role of the Central Institute of Statistics.

The experience gained from sample techniques with the aim to study the labor force and the labor market, provided methodological tools and impetus for the use of such techniques for the acquisition of information in other social areas. In presenting the supplement of the Italian Education Statistical Yearbook of 1955, dedicated to the "special survey on college students and graduates of high schools," Maroi wrote: "I am confident that the present research, which corresponds to the Institute's aim of broadening and deepening evermore every field of inquiry, is able to meet the expectations of scholars," with the explicit vision of what he envisioned, which introduced the concept of several users of statistics apart from the political decision-makers.

Combined with the labor force survey of 1957, special research was also completed, published in 1958, "on some aspects of living conditions of the population." The survey indicates how the perspective of the social factor of official statistics was expanding. Furthermore, the focus on the education system and labor force, a priority until then, outlined the potentialities on which the political powers could count on in order to define development policies. Surveys that touched on individual choices and lifestyles outlined a first approach to the study of quality of life by looking into what people read, the use of new technologies of the time—radio, television, cinema—smoking habits, changes, and aspirations of working fathers and sons, the prospects for social mobility for themselves, and for future generations. The respondent was the householder, who provided the information for all other components, and so the variables used were only referred to the head of the family. Such variables were interpreted as different variables of the behavior related to reading without taking into account how the higher education in the younger generations could also modify the social habits of the whole family. Yet combined with the labor force surveys of 1965 and 1973, the next two surveys were carried out specifically focused on the reading habits of individual members of the family in relation to newspapers, periodicals, and books not read for purposes of study and work. The three surveys, repeated 7 or 8 years apart, foretell a new direction to build on in areas more closely related to social issues, not just demographic or economic, of the longitudinal paths, with reading being able to highlight and monitor the resulting changes of behaviors and lifestyles.

The two censuses that fall in this period testify to the improvement in living conditions. In fact, the questions on housing conditions in 1951 are extended to also consider the presence of a heating system and, in 1961, also the type of system be it centralized or autonomous as well as specifying between services inside or outside the home (1951) and between the toilet and bathroom (1961): this ranges from measuring the satisfaction of needs to the level of comfort. In this light, beginning from 1971, the time taken to travel to the place of study or work, and the means of transport are collected. In the 1961 and 1971 censuses, there is a section dedicated to the marital and reproductive life of women, which subsequently will not be replicated.

4 The Push Toward International Comparable Statistics and Social Indicators

Once again, a shock to the implementation and modernization of statistics comes from the cultural movements and international institutional arrangements of new supranational institutions. Social statistics in particular hold a new and stronger standing along with attention to which countries look to social inequality, belonging to social classes, lifestyle, and marginality. The social indicator movement, born in the US, had attracted the interest of all other Western countries, including Italy. In 1971 the XXVII Scientific meeting of the Italian Society of Statistics dedicated its work to social indicators. The scholars of the SIS, then as now, were mostly academics but the presence of members of ISTAT and central government, whose contributions would change the perspective of the production of official statistics in the social field, was significant. The birth of the OCSE in 1960 intensifies the pressure to produce statistics suitable to compare the political experience and living conditions of the member countries.

The quarterly surveys on labor force had become fully operational. The yearbook of statistics on education was systematically published using administrative data implemented from time to time by special surveys. From a thematic dimension other social issues present in the Italian Yearbook of Statistics become more detailed. The systematic production is increasingly complemented by occasional surveys, and referenced to specific social issues (the survey conducted on holidays every year since 1959, and the survey on sport).

But a more systematic reflection on social statistics appeared in the Second Conference on Information Statistics held in 1981 (Golini 1981, Rey 1981), where an entire session was devoted to social statistics and which formed the basis for the publication of the second volume of social statistics—the first had appeared in 1975 (Istat 1975, Moser 1983). In the next volume of 1993 the title would change to "Statistics and social indicators." In the presentation Rey, the ISTAT president at the time clarified the choice, "Without neglecting the absolute values, special attention was devoted to the preparation of reports and indices, which allow an immediate comparison between the configurations that individuals take in the various phenomena regions. To this end were used, where possible, coded and characteristic ratios now commonly used, while in other cases less common solutions have been proposed, notwithstanding the immediate preservation of the need for comprehensibility and relevance of the developed measures (Istat 1993)." This approach represents a significant change in the role of official statistics in the direction of facilitating the interpretation of the information produced (Fiocco 2009).

5 Household Surveys

From this gradual growth of attention to social issues comes the "Survey on family structures and behavior (Istat 2009)." This was completed for the first time in 1983 in order to deepen the study of the family structure along with family relations and the system of free assistance. This survey is a prerequisite to the Multipurpose Survey on families which kicks off the project in its earliest form in 1987 to complete the planned cycles (excluding the final that would never be realized) in 1991. In the first round the survey collected information on new family forms (such as free unions) and the life-cycle of women, with an emphasis on increased fertility and marriage histories. The IMF is a fixed point and totally innovative for social statistics in Italy. Firstly it was intended with the declared aim of summarizing in a single design, in order to detect all of the social issues of interest so as to be able to compare, both transverse and longitudinal, the social dimensions of people's living conditions and their transformation over time. This was achieved by monitoring changes themselves and in relation to changes that were occurring in other dimensions or within the family structure. It also extends many areas of interest defined in the 1981 volume of the Social Statistics through the introduction, from time to time and in different cycles, of: the victimization by crime, home accidents, conditions of disability, use of time, short and long-term travel, school activities and conditions of childhood, the condition of the elderly, family networks, use of social and health services and hospital use of medications, chronic diseases, and smoking habits. All these issues are now dealt with by the individual and not on the part of the institutions. Hence by the demand rather than by the offer, completing the information of an administrative nature might come from the school system, the health system, the social security system, the justice system, and so on. In this way, comparisons and differences were highlighted between statistics from administrative sources and the information provided by individuals. This is especially relevant when considering the uncertain number of unreported crimes, domestic accidents that result in hospitalization or absence from work in health statistics, school dropouts not due to official withdrawal from courses, living conditions, and care of the disabled, assistance and welfare statistics.

Between June 1988 and May 1989, the first national survey on the use of time took place, indicating the different life styles and behavior in time management between different social subjects according to gender, age, family structure, and so on. Taking the side of the individual also means following the path of detecting the subjective dimension. Thus a subjective and satisfaction approach will increasingly be used in subsequent surveys by the IMF survey system. The "pillar" survey of the current IMF is on "Aspects of Daily Life," conducted annually since 1993, when the system had been redesigned. It collects, in fact, all the phenomena which are then detailed in thematic research (Gazzelloni 2004).

The survey "Citizens and leisure" was founded in 1995 as an attempt by ISTAT to systematically describe a sector which is very tied to choices and subjective perceptions, such as leisure and relationships between the latter and cultural participation.

The survey on public safety, also called on the victimization survey and carried out for the first time in 1997, takes the side of those who have suffered a crime, even if not reported. The characteristics of the victims are put in evidence and the types of weaker subjects or those more easily attacked by specific crimes emerge. In addition it is the first survey in 2006 on violence against women in a framework of collaboration with the Department for equal opportunities.

Another international boost comes in relation to time use surveys that were conducted in several countries with different methodologies. In Italy the first survey along the lines proposed by the Statistical Program Committee was carried out in 2002/2003 and the second survey in 2008/2009. The time use surveys provide a comprehensive and effective tool for delineating the lifestyles, conditioning, and behavioral choices of individuals with regard to age, gender, and family structure. Furthermore in 1998 the first survey "Family, social subjects and conditions of childhood" was conducted and repeated in 2003 with the same structure but with renovated and expanded content. In this survey, emergent phenomena such as prolonged permanence at home by young people in the family of origin and the postponement of marriage and reproductive projects by women would be studied.

An important social survey, created under an agreement between ISTAT and the Ministry of Labour and Social Policy, is on social integration of persons with disabilities through which the factors that hinder the full participation to economic life and social development of the country are analyzed. The survey, conducted between January and March 2004, was addressed to persons who were disabled or impaired in activity at the time of the survey "Health status and use of health services" in 1999/2000 and is therefore a further concrete example of integration and synergy between the various surveys of the system.

Another important survey to define the living conditions of households is on consumption: this survey, although falling within conceptual economic statistics, highlights the different lifestyles and monitors changes in dietary and spending behavior according to family structure and class. The information on household consumption over time becomes an information base for the work of the commission of inquiry on poverty and to define the poverty line in Italy in both absolute and relative terms. This approach marks the final introduction of complex indicators into official social statistics (Facioni 2004).

6 The Role of the Birth of SISTAN Within the Change in Approach to Social Statistics

The IMF experience finds its final structure during another event that led to a radical review of the Italian statistical system: the establishment of SISTAN in 1989. This event defines the process of transformation of ISTAT into a research institution and this step will bring about a number of consequences including the effort to provide users not only with data but also methods of analysis. From the new structure deriving from SISTAN, a set of tools and activities which reorganize the production of official statistics in the field of social statistics took

place. The direct involvement and accountability of government institutions and activities in SISTAN release ISTAT from the burden of a systematic series of surveys deriving from administrative tasks and assigned, from time to time to the competent institutions that take responsibility for making available and publishing objective data related to their institutional activities. Alongside the above mentioned, also specific research interests which produce further statistic information that is fully available to the users and interacting with those produced by other actors within the system, emerge. In particular, the surveys carried out in collaboration between ISTAT and ISFOL deepen social issues on work, on inequality in general, and on unemployment. The survey of the BI on Italian household budgets, carried out since the 1960s, is completely redesigned to provide data integrity with the ISTAT survey on consumption.

With the birth of SISTAN, the real activities of ISTAT research with *ad hoc* surveys carried out on their own, or in conjunction with other ongoing activities of other institutions, expand. The integrated system of surveys on the education-work transition, which began as early as 1989 and aimed at the analysis of subsequent pathways to achieving the graduation of young people, who successfully conclude a course of post-compulsory study, is carried out every 3 years and after the third year after graduation and is part of the first type of research. The survey on separation and divorce, conducted yearly by ISTAT in the civil courts since 1969 for separations and divorces (from 1971 for divorces and the survey on child custody) belongs to the second type. But today the substantial legislative changes, both in marriage break ups and child custody, make these surveys strongly explanatory of change in behaviors and attitudes toward value-pairs of the traditional family. Another type of social investigation that relies on the collaboration with administrative tasks is related to applications for adoption. The Invalsi activity, whose statistics on the learning of children at various levels of education should also be mentioned.

7 The Surveys on Living Conditions of Families in Europe: EU-SILC and the New ISTAT Social Surveys Recently Implemented or Being Planned

Since 2004, Eurostat and the NSI of Europe made available to scholars and policy-makers a broad range of information on living conditions of European households drawn from sample surveys and administrative sources. But the most comprehensive survey of European social content is the EU-SILC survey carried out under the European regulation in order to allow comparability between the data collected from member countries. This survey is particularly interesting since it is fully designed to meet the needs of monitoring of the Lisbon strategy and already provides as output the Laeken indicators, calculated and compared.

In the same year the new Labour Force Survey started; the change was undertaken in line with European Union regulations. A significant feature of the survey is the establishment of new criteria for identifying employed and unemployed individuals, as well as a far-reaching reorganization of the data collection and production process.

In 2006 a new survey for the first time was entirely devoted to the phenomenon of physical and sexual violence against women while surveys on sexual harassment and violence were conducted in 1997 and again in 2002 as part of the Multipurpose Survey "Safety of citizens."

The current interest in ISTAT social activities is further evidenced by the new special surveys recently completed (the first survey on "Income and living conditions of the foreign population resident in Italy" conducted on a sample of 6,000 families with at least one foreign member), or soon to be realized (statistics of homeless people and those who do not occupy a house, that once completed, will highlight the emerging and upcoming phenomena such as new types of poverty and marginalization and the Adult Education Survey that aims to detect the involvement of adults in training).

8 Conclusions

The widening scope of social statistics that occurred in the production process does not seem destined to stop. Indeed, two forces of different nature, in our opinion, will mark the future of social statistics: the local system of public management, partly as a result of orientation to federalism, is burdened with tasks that will require statistics at a greater level of disaggregation in terms of both territorial and individuals and social groups' segmentation. On the other hand, compared to local requirements, there are many pressures that come from supranational institutions toward methodological and informational solutions with the purposes of knowledge of social conditions of countries as a whole. *In primis*:

- The OECD project "Measuring the Progress of Society" (to support national initiatives for the definition of well-being and progress).
- Communication from the European Union Commission to the Council and the European Parliament "GDP: measuring progress in a changing world" (the EU guidelines to complement GDP with social and environmental indicators).
- The final report of the Stiglitz Commission (with recommendations to improve the measurement of economic performance, quality of life, and environmental sustainability).
- The "Sofia memorandum," signed September 30, 2010 during the 96th Conference of Presidents and General Managers of NSI in Europe (the issue of measuring progress, prosperity, and sustainable development is a key element of official statistics).

From these new forms of measurement of progress and prosperity will come certainly the need for social statistics information more disaggregated and richer than those available today.

References

Cortese, A.: I censimenti della popolazione. In: Cinquanta anni di attività. 1926–1976. Istituto centrale di statistica, pp. 89–98, Roma (1978)

Facioni, C.: Breve storia dell'evoluzione delle statistiche sociali e di genere nell'ambito delle statistiche ufficiali dell'ISTAT. In: Fraire, M. (ed.) I bilanci del tempo e le indagini sull'uso del tempo. CISU, Roma (2004)

Fiocco, B.: Le misure dell'Italia nell'Annuario Statistico italiano, Documenti ISTAT n. 3 (2009)

Gazzelloni, S.: Il sistema di indagini multiscopo sulle famiglie dal 1993 a oggi. In: Fraire, M. (ed.) I bilanci del tempo e le indagini sull'uso del tempo, CISU, Roma (2004)

Golini, A.: Le statistiche sociali, Annali di Statistica, Serie IX – vol. 1 (1981)

Istituto centrale di statistica, Commissione Alleata e Presidenza del Consiglio dei Ministri. Censimenti e Indagini per la ricostruzione nazionale eseguiti nel settembre 1944. pp. 51–68, Roma (1945)

Istituto centrale di statistica. Le rilevazioni statistiche in Italia dal 1861 al 1956. Censimenti della popolazione e delle abitazioni, Annali di statistica, serie VIII, vol. 5, pp. 347–386, Roma (1957)

Istituto centrale di statistica. Le rilevazioni statistiche in Italia dal 1861 al 1956. Modelli di rilevazione. 1. Censimenti – Statistiche demografiche e sociali. Annali di statistica, serie VIII, vol. 8, pp. 4–70, Roma (1959)

Istituto centrale di statistica. Statistiche sociali, Annali di Statistica, Serie IX, vol. I, Roma (1975)

Istituto centrale di statistica Statistiche sociali, Annali di Statistica, vol. II, Roma (1981)

Istituto centrale di statistica: Statistiche e indicatori sociali. Annali di Statistica, Roma (1993)

ISTAT – Istituto nazionale di statistica. Navigando tra le fonti demografiche e sociali. Istat, Roma (2009)

Rey, G.: Orientamenti di una politica per la statistica negli anni '80, Annali di Statistica, Serie X vol. 3, pp. 1–16 (1981)

Rapporto Moser. Aspetti delle statistiche ufficiali italiane. Esame e proposte, Annali di Statistica, pp. 26–40 (1983)

Response Burden Reduction Through the Use of Administrative Data and Robust Sampling

Maria Caterina Bramati

Abstract

There are several reasons why robust regression techniques are useful tools in sampling design. First of all, when stratified samples are considered, one needs to deal with three main issues: the sample size, the strata bounds determination and the sample allocation in the strata. Since the target variable y, objective of the survey, is unknown, it is used some auxiliary information x known for the entire population from which the sample is drawn. Such information is helpful as it is strongly correlated with the target y, but of course some discrepancies between them may arise. The use of auxiliary information, combined with the choice of the appropriate statistical model to estimate the relationship with the variable of interest y, is crucial for the determination of the strata bounds, the size of the sample and the sampling rates according to a chosen precision level of the estimates, as it has been shown by Rivest (2002). Nevertheless, this regression-based approach is highly sensitive to the presence of contaminated data. Indeed, the influence of outlying observations in both y and x has an explosive impact on the variances with the effect of strong departures from the optimum sample allocation. Therefore, we expect increasing sample sizes in the strata, wrong allocation of sampling units in the strata and some errors in the strata bounds determination. Since the key tool for stratified sampling is the measure of scale of y conditional to the knowledge of some auxiliary x, a robust approach based on S-estimator of regression is proposed in this paper. The aim is to allow for robust sample size and strata bounds determination, together with the optimal sample allocation. To show the advantages of the proposed method, an empirical illustration is provided for Belgian business surveys in the sector of Construction.

M.C. Bramati (✉)
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza, University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy
e-mail: mariacaterina.bramati@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_12,

It is considered a skewed population framework, which is typical for businesses, with a stratified design with one *take-all* stratum and $L - 1$ strata. Simulation results are also provided.

1 Introduction

The presence of outliers can strongly bias the sampling design and hence the survey results. In particular, it could induce a wrong computation of the number of statistical units to sample, usually overestimating it.

In what follows we focus on the stratified sampling design, which has been proven to be the most efficient surveying technique under some basic assumptions (see Tillé 2001) and it is currently in use at several NSIs for business surveys.

For instance, suppose that in the stratification variable X some outliers arise. Outliers are observations arbitrarily far from the majority of the data. They are often due to mistakes, like editing, measurement and observational errors. Intuitively, when outliers are present in a given stratum for the stratification variable X they affect both the location and scale measures for X. Therefore, it is clear that a higher dispersion than the "true" one will be observed in that stratum.

Such a situation will bias the outcome of the HL method. For instance, the sample size would be bigger than it should be, given the fact that observations seem to be more distant (in average) than they are in the reality. Moreover, the strata bounds and the sample allocation would be both biased. This is clear when we consider the Neyman allocation, for example, which is based on within-stratum dispersion. Since the principle is to survey more units in the strata in which the auxiliary variable is more dispersed within the stratum, outliers might have the effect of increasing enormously and unduly the sample size in each stratum.

For this reason we build two robust versions of the HL method, the *naive robust* and the *robust* HL sampling strategy which we compare through a simulation study.

2 The Problem

We focus on simple stratified samples with one take-all stratum and several take-some strata. This because we deal with

- skewed distributions (small number of units accounts for a large share of the study variables)
- availability of administrative information, providing a list of the statistical units of the target population (i.e. tax declaration, social security registers)
- survey burdens for firms and costs for NSIs
- data quality (administrative sources and survey collection)
- compliance requirements established by EUROSTAT

Now, it is known that there exists a discrepancy between the auxiliary variable X used for stratification and the survey variable Y. Therefore, the strategy suggested by Rivest (2002) is to recover such discrepancy by the use of a regression model.

Of course, the auxiliary information X is only a proxy for the target variable Y, which requires to estimate the *discrepancy* between Y and X, as suggested by Rivest (2002) with the *modified* HL algorithm.

In the business survey literature, the relationship existing between Y and X is often modeled by a log-linear regression relationship. Let X and Y be continuous random variables and $f(x)$, $x \in \mathbb{R}$ the density of X. The data $x_1, \ldots, x_N$ are considered as N independent realizations of the random variable X.

Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h, for $h = 1, \ldots, L-1$.

This model considers the regression relationship between Y and X expressed by

$$\log Y = \alpha + \beta_{\log} \log X + \varepsilon,$$

where ε is assumed to be a 0-mean random variable, normally distributed with variance $\sigma^2_{\log}$ and independent from X, whereas α and $\beta_{\log}$ are the parameters to be estimated.

However this approach presents some weaknesses

1. s^2_{yh} is unknown, which makes crucial the use of the auxiliary information X
2. the number L of strata is selected by the user
3. the administrative records are often of low quality (errors)

We can distinguish three main sources of anomalies, listed below

- erroneous records in the surveyed data (Y) (**vertical outliers**)
- quality issues in the administrative registers (X) (**leverage**)
- outliers in both variables (X, Y) (good/bad **leverages**)

The presence of such anomalies makes unreliable the conditional mean and variance of $Y|X$, therefore affecting the sample size and strata bounds determination as well as the sample allocation.

In what follows we propose a possible alternatives to the Rivest (2002) modified HL algorithm. Strata bounds and sizes are derived minimizing the conditional variance in each stratum after a re-weighting of the information according to the degree of *outlyingness*. We refer to this approach as to the *robust modified HL* algorithm.

3 The Robust Modified HL Algorithm

Supposing that a log-linear relationship exists between the survey variable Y and the auxiliary one X, then consider the S-estimator of regression as in Rousseeuw and Yohai (1984) as

$$S(x, y) = \arg\min_{\beta} s(r_1(\beta), \ldots, r_N(\beta))$$

where $r_i(\beta)$ are the regressions residuals and s is scale measure which solves

$$\frac{1}{N}\sum_{i=1}^{N}\rho\left(\frac{r_i(\beta)}{s}\right)=b$$

for a conveniently chosen ρ function and a constant b. This estimator is robust with respect to both vertical outliers and leverage points. Then, with some straightforward calculations (expanding $\rho(\cdot)$), the following approximation holds

$$\mathrm{Var}[Y\,|b_h \geq X > b_{h-1}] \approx e^{\sigma^2}\psi_h/W_h - (\phi_h/W_h)^2,$$

where

$$W_h = \int_{b_{h-1}}^{b_h} \omega(x^\beta) f(x)dx \tag{1}$$

$$\phi_h = \int_{b_{h-1}}^{b_h} x^\beta \omega(x^\beta) f(x)dx \tag{2}$$

$$\psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} \omega(x^\beta) f(x)dx, \tag{3}$$

β and σ are the parameters of the log-linear model in the previous section, and $\omega(x) = \rho'(x)/x$ is the weighting function.

The problem then reduces to solving for bounds $b_1, \ldots, b_h, \ldots, b_L$ which minimize n using the Neyman allocation scheme. In symbols, under the loglinear specification the objective function is

$$n_{\hat{\iota}_{\text{ystrat}}} = N_L + \frac{(\sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2})^2}{(c\sum x_i^\beta/N)^2 + \sum_{h=1}^{L-1}\frac{(e^{\sigma^2}\psi_h - \phi_h^2/W_h)}{N}} \tag{4}$$

where *robust* moments W_h, ϕ_h and ψ_h are those in (3), β and σ are the parameters of the log-linear model estimated by robust regression (S-estimator or LTS).

Then, the Sethi's iterations are run for a given L and precision c, computing the optimal strata bounds and sample size.

4 Simulation Study

The aim of the simulation study is to compare the performance of the two robust sampling strategies proposed in this paper with respect to Rivest (2002)'s based on classical LS regression.

Table 1 Summary of results comparing Robust modified HL method versus modified HL (Rivest 2002), target precision: 1 %

Design	Relative efficiency	Relative sample size
No outliers	0.10	100
Long-tailed Cauchy	0.00	0.29
Long-tailed t	0.08	10
Vertical outliers 15 %	0.99	10
Leverage points 15 %	0.00	10
Vertical outliers 30 %	0.99	1.43
Leverage points 30 %	0.00	1.43

Simulations are performed using the business sampling frame of the Structural Business Survey in 2002, where we consider as target variable (y) the value added of enterprises in the industry of *Constructions* which are stratified by the economic-size class. The number strata $h = 1, \ldots, 6$ is set according to the common practice in SBS, with 1 take-all stratum and 5 take-some strata. The auxiliary information x is on the turnover (from the VAT register). Then, population is generated from

$$\log y_i = \beta \log x_i + \varepsilon_i$$

with a choice of $\beta = 0.75$.

Then we consider the following designs

1. no outliers: $\varepsilon_i \sim \mathcal{N}(0, 1)$
2. long-tailed errors: $\varepsilon_i \sim \text{Cauchy}_1$
3. long-tailed errors: $\varepsilon_i \sim \text{t}_3$
4. vertical outliers: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}\left(5\sqrt{\chi^2_{1;0.99}}, 1.5\right)$
5. bad leverage points: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}(10, 10)$ and corresponding $X \sim \mathcal{N}(-10, 10)$.

The contamination level, i.e. the percentage of outliers in the data, is set to $\delta = 15$ and 30 %. Then the three procedures are used to compute the strata bounds, sizes and allocation

- generalized HL method (Rivest 2002)
- robust generalized HL method

at 1 % precision and compared by means of relative MSE of the Horvitz–Thompson estimator for the mean. In Table 1 are displayed the main results.

References

Rivest, L.P.: A generalization of Lavallée and Hidiroglou algorithm for stratification in business surveys. Techniques d'enquêtes **28**, 207–214 (2002)

Rousseeuw, P.J., Yohai, V.J.: Robust regression by means of S-estimators. In: Franke, J., Hardle, W., Martin Robust, D. (eds.) Nonlinear Time Series. Lecture Notes in Statistics. vol. 26, pp. 256–272. Springer, Berlin (1984)

Tillé, Y.: Théorie des sondages. Dunod, Paris (2001)

An Application of Text Mining Technique for the Census of Nonprofit Institutions

Domenica Fioredistella Iezzi, Massimo Lori, Franco Lorenzini, Manuela Nicosia, and Sabrina Stoppiello

Abstract

The National Institutes of Statistics are increasing in the use of administrative data, which are routinely collected by organizations as part of their business or operational activities. As a matter of fact, this huge amount of data is relevant whether by transforming in statistics for building information systems or by using them as additional information during the whole statistical survey. With regard to Italian nonprofit institution^{tm} Census, text data from the Italian Revenue Agency are being used to create the list. The paper explores the opportunity of using the text mining technique on the available data to build a classification of nonprofit organizations, which may also help to distinguish them from firms and public institutions. We apply text classification and text clustering methods to select the best partition of the dataset, and we underline the advantages and disadvantages of different processes.

1 Introduction

General Censuses of Industry and Services were carried out in 1981, 1991, and 2001 (Istat 2001). The National Institute of Statistics (Istat) has been surveying nonprofit organizations (NPI) as well as firms and public institutions. With regard to nonprofit sector, one of the most critical aspects in the Census is the level of coverage due to the lack of comprehensive and base archive for NPI. In fact, the identification of the

D.F. Iezzi (✉)
Tor Vergata University, Rome, Italy
e-mail: stella.iezzi@uniroma2.it

M. Lori • F. Lorenzini • M. Nicosia • S. Stoppiello
DCCG, ISTAT, via A. Ravà, 150, Rome
e-mail: malori@istat.it; lorenzin@istat.it; mnicosia@istat.it; stoppiel@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_13,

NPI is not an easy task, due to the lack of a legal framework, as well as a not clear and general definition of nonprofit institutions into the Italian law. Furthermore, NPI are often informal organizations. Nevertheless, there are different sectorial archives of NPI (Voluntary Organizations and Social Cooperatives Registry, EAS, 5 per mille, Onlus, etc.), but these registries do not cover the population of NPI. Since 1999 Istat has consulted the Italian fiscal register (by Agenzia dell'Entrate) avowing to consider firms and public institutions in order to ensure high level of coverage. In the context of the Census it is essential to identify NPI within units (over three million) registered into the fiscal archive differentiated by legal profiles both public and profit organizations. Although the fiscal registry classifies the units by legal framework and economic sectors, the recognition of nonprofit status is not prompt. Indeed, some preliminary analyses have shown that the fiscal registry classification by legal status and economic sector may be biased. After all, administrative data do not often satisfy the objectives and the criteria of the statistical surveys.

In this paper, we propose two approaches to classify NPI: (1) text categorization to build automatically by learning category properties from a set of pre-classified NPI; (2) text clustering to detect the best partition, using Affinity Propagation (AP) and Partitioning Around (PAM) algorithms.

The paper is organized as follows: in Sect. 2, we present method 1; in Sect. 3, we illustrate method 2; in Sect. 4, we describe the case study and the main results and, finally, in Sect. 5, we expose the conclusions and the future developments.

2 Text Mining and Text Categorization

With regard to the NPI identification, it is useful to consider the organization name recorded into the fiscal registry. Furthermore, it is possible to classify NPI by institutional typology capturing the semantic meaning of the corporate denomination. To analyze the content of organizations name, it is necessary to apply text mining techniques. According to literature, text mining implies a wide range of methods used to select the targeted information in a large number of documents and it automatically identifies interesting patterns and relations in textual data (Bolasco 2005). In detail, the content analysis of fiscal registry unit name could be considered as case in point of text classification. Starting from a textual document set and a taxonomy, the overall objective of text categorization is a process to identify the proper category for each document. In practice, the aim is to find a function M: $D \times C \rightarrow \{0, 1\}$, called classifier, where D is a set of disordered textual data and C represents a set of the categories of a taxonomy. The M function has to be formulated so as its behavior is as closer as possible to the true (but unknown) assignment function of textual document to the appropriate category. Generally speaking, text classification techniques could be divided into two main branches. The first is the knowledge engineering approach which focuses on manual development of classification rules. A sectorial expert identifies a set of sufficient conditions for a document to be categorized into a given category. For each category the classification process consists of logical rules showed as follows:

if DFN formula, then category=c_i. The system DNF (disjunction of conjunctive clauses) is simply a list of logical clauses, therefore, if a document satisfies a specific condition, that is classified into the related category (Sebastiani 2002).

On the other hand, the machine learning (ML) approach whereby the classifier is built automatically by learning category properties from a set of pre-classified examples. In the ML approach, there are many techniques to build the classifier learning; some of them are based on probabilistic theory or decision tree (Yang 2006), and others use neural network algorithms (Bishop 1995).

While the main disadvantage of knowledge engineering approach is the huge amount of time and expert knowledge required, machine learning approach is much less costly to implement but generally it does not reach the performance in comparison with the other one (Feldman and Sanger 2006).

3 Text Clustering

Text clustering (TC) is a process of assigning documents to clusters based on the characteristics they possess and it is often the only available solution to organize large collections of text on different topics.

We use TC to classify the names of the NPI in a way that each expression is in the same cluster if it is very close to other words of the same cluster, with respect to some criteria of similarity. This clustering process allows to detect automatically the class of belonging of an institution.

Let a corpus C of NPI, represented by a vector of weighed terms of the form: $dj = (w_{1j}, w_{2j}, , w_{ij}, , w_{pn})$, where w_{ij} is the weight for term i, attached to NPI d_j. By joining these vectors, we get the **D** word-term-by-document-matrix. We use term frequency scheme: $w_{ij} = n_{ij}$, where n_{ij} is the number of NPI in which term j occurs (Iezzi 2012a). Each NPI is described by a vector of term frequencies, the well-known bag-of-words model. We calculate cosine distance on matrix **D** to measure the similarity between the denomination of NPI. The cosine similarity takes into account only the angle and discards the magnitude (Iezzi 2010). It measures the cosine of angle formed by two document vectors that describe two NPI, A and B. Formally, the cosine distance is:

$$\cos(\alpha) = \frac{< A, B >}{|A|B|}$$

We use the cosine distance as an input for Multidimensional Scaling (Proxcal procedure) for exploring similarity and dissimilarity in data (Borg and Groenen 2005). We apply two algorithms (AP and PAM) to select the centroids of the clusters from the denominations of the NPI and compare the results. We want to detect centrotypes, that represent the prototypes of the group, and both AP and PAM select centroids within of the input matrix.

AP simultaneously considers all keywords, selected by a list, as potential exemplars or centers (Frey and Dueck 2007). Exemplar-based clustering that is the

Table 1 Lexical measures of the corpus

Word tokens (N)	2,025,373
Word types (V)	134,554
Type/token ratio (TTR)	6.643
Hapax number (V)	82,017
Hapax percentage (V1/V)	60.955%
Average frequency (N/V)	15.052

task of not only performing the partitioning but also identifying for each cluster its most representative member, or exemplar. A common characterization for the cluster exemplar is the data point whose overall similarity to other data points in the cluster is maximal. By viewing each data point as a node in a network (Iezzi 2010, 2012b), this method recursively transmits real-valued denominators along edges of network until a good set of exemplars and corresponding clusters emerges. The input of AP is the pair-wise cosine similarities between each pair of key names of NPI, $c[i, j](i, j = 1, 2, \ldots, N)$. Given similarity matrix $\mathbf{C} = [ij]$, AP attempts to find the exemplars that maximize the net similarity, i.e. the overall sum of similarities between all exemplars and their member data points. PAM algorithm is a classical partitioning technique of clustering, in which k clusters are known a priori. To evaluate the PAM results and to determine k, we use the silhouette index (Rousseeeuw 1987). We prefer PAM instead of the well-known k-means algorithm, because it is more robust to noise outliers. Moreover it selects medoids, that represent the most centrally located point in the cluster (Iezzi 2012b). At the end we compare the AP and PAM results and evaluate the best algorithm in relation to the ability to interpreter.

4 The ISTAT DB on Non-Profit Institutions

The program, more than 1,000 classification rules, was tested on the dataset of 435 thousand units recorded in the fiscal registry, presenting theoretically a legal framework consistent with the status of NPI.[1] The corpus is composed of 2,025,373 word tokens (N), and 134,554 word types (V) and 60.9 % of total forms are hapax (Table 1).

The lexical measures of the corpus show some weakness from the statistical point of view. It is, in fact, a corpus with a high percentage of hapax (about 61 %), which it is far from the threshold recommended for quantitative analysis, accounting for 50 %. The type token ratio is equal about to 7 %; this value is more comforting, in fact, the threshold is less than the recommended 20 %. Based on these parameters, we can conduct a quantitative analysis focused largely on descriptive aspects, relying on the classification and concordances of words.

[1]The number of total words is equal to 1,719,774 while the number of graphical forms is 45,413. The percentage of hapax is around 35.1 %.

Table 2 More frequent full words in the ISTAT DB

Word types	No.
Associazione (Association)	84,618
Condominio (Condominium)	53,601
Via (Street)	23,098
Parrocchiale (Parish)	22,541
Culturale (Cultural)	19,205
Club (Club)	19,189
Sportiva (Sports)	16,615
Comitato (Committee)	16,324
San (San)	15,345
Circolo (Circle)	14,511
Beneficio (Benefit)	14,414

Table 3 More frequent full repeated segments in the ISTAT DB

Repeated segments	No.
Associazione culturale (Cultural association)	15,746
Associazione sportiva (Sports association)	14,297
Beneficio parrocchiale (Benefit parish)	11,177
Associazione sportiva dilettantistica (Amateur sports association)	8,473
Chiesa parrocchiale (Parish church)	7,671

The more frequent word is "associazione" (association), that is presented 84,618 times; followed by "condomio" (condominium), that appears 53,601 times and "parrocchiale" (parish) with 22,541 occurrences (Table 2).

More frequent full repeated segments are "associazione culturale" (cultural association), that appears 15,746 times; "associazioni sportive" (sporting association) with 14,297 occurrences; "beneficio parrocchiale" (parish benefice) with 11,177 occurrences, "associazione sportiva dilettantistica" (Amateur Sports Association) with 8,473 occurrences, and "chiesa parrochiale" (parish church) with 7,671 occurrences (Table 3).

4.1 Text Categorization Results

To classify the different units from both firms and public institutions, recorded in the fiscal registry, we adopted the knowledge engineering approach. This problem is developed as a research of target entity (in this case, considering different categories of NPI) that is scattered inside a collection of texts. Figure 1 shows that the repeated segment "ASSOCIAZIONE SPORTIVA" (sport association) is closely related to organizational framework (society, club, center, . . .), and several attributes of nonprofit institutions (football, horse, bowling, sailing, tennis, . . .).

Back to back mild pre-treatment of the corpus removing the stop-words (mainly articles and preposition), the phase of identifying the specific dictionary and recurrent segments for each category of the institutional typology is needed. To this

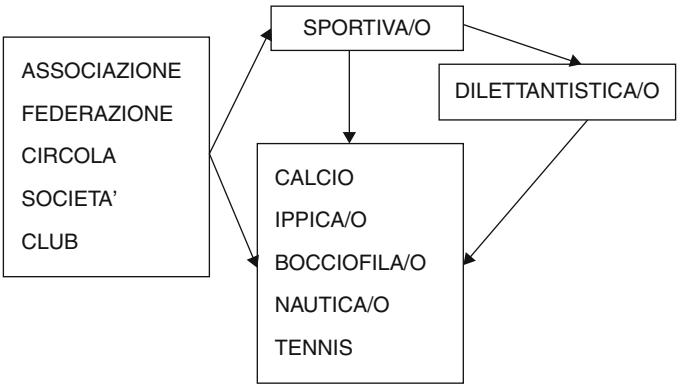

Fig. 1 Research of the repeated segment "associatione sportiva"

purpose, the software Lexico 3 was used as training to set the organizations names in 1999 Census. The automatic classification program related to the organization's name was written in SAS as character functions of this software (scan, substr and find) in order to locate categories' distinctive attributes into a text string. In this way, the position of the words is deeply related to the assignment of an organization and the fitting category. Altogether, this procedure has produced good results classifying about 75 % of units; in addition, about 50 % of classified records have been recognized as "association" coherently with the distribution of Italian nonprofit institutions differentiated by legal framework. Lastly, about 30 % of organizations do not match the criteria of NPI (firms, public institutions, condominia, etc.). The assessment in depth of the performance of the adopted procedure was important to compute two measures: recall errors and precision errors.[2] For this purpose, a set of 450 units were casually sampled from the original dataset. Hence, 60.4 % and 79.1 % are, respectively, recall and precision errors and 68.5 % is the combined indicator (F) for the abovementioned errors. The F is the harmonic mean of precision and recall, and it assumes the following formula:

$$F = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision is the fraction of retrieved documents that are relevant to the search, and recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

[2]The precision error occurs when a document is not assigned to the pertinent category by the classifier while recall error happens when a category does not include documents that should belong to it. The first type of error is measured as the percentage of correctly classified documents among those attributes to the category and the latter is defined as the percentage of correctly classified documents among all those belonging to that category.

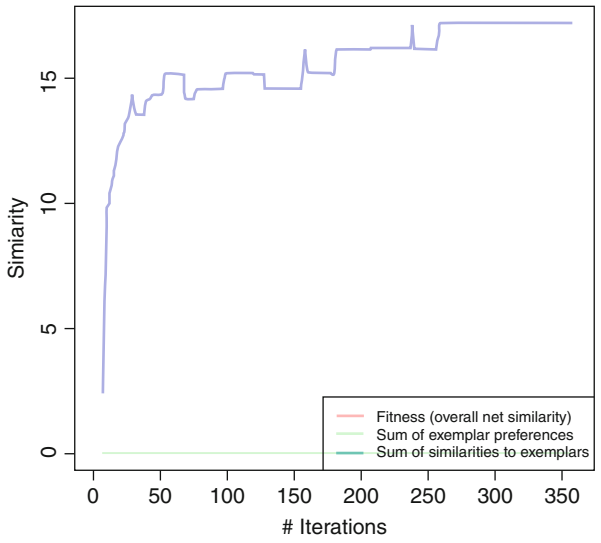

Fig. 2 Performance measures using AP

4.2 Text Clustering Results

We studied the problem to classify the organization name recorded into the fiscal registry using the standard optimization criterion of squared error. Both AP and PAM algorithms (Kaufman and Rousseeuw 1990; Iezzi 2012b; Iezzi et al. 2012) are used to categorize 189 key denominations of 434,035 NPI. AP found the best solutions after 334 numbers of iterations, it selected 1,719,188 net similarities and it identifies the best classification into 38 clusters (Fig. 2). Figure 2 shows the three performance measures that AP uses internally for each iteration: (1) Sum of exemplar preferences; (2) Sum of similarities of exemplars to their cluster members, (3) Net fitness: sum of the two former.

Table 4 shows that, using silhouette index (SI), the best partition is composed of 35 classes, because SI is the.

AP algorithm classifies very well several groups, e.g. the exemplar "genitori" (parents—cluster g17) is the center of the class that contains "infanzia" (childhood), "liceo" (secondary school), and "scuola" (school) or the exemplar "fitness" to which belong expressions "palestra" (gym), and "riabilitativo" (rehabilitative) . The advantage of the AP method is to find a quick solution, the disadvantage is that creates more classes composed of a single unit or mixed classes. The clusters g1, g2, g3, and g4, e.g., hold only the exemplars; the cluster g6, detected by the exemplar "association," incorporates many different keywords, e.g. "accademia"(academy), "amici" (friends), "amicizia" (friendship), "teatro" (theater), creating a not well-defined class (Fig. 3).

Table 4 Silhouette index from 30 to 40 clusters using PAM algorithm

k	30	31	32	33	34	35	36	37	38	39	40
PAM	0.50	0.55	0.56	0.58	0.61	**0.63**	0.62	0.59	0.58	0.59	0.57

CLUSTER	EXEMPLAR	KEYWORDS
g1	AMATORI	
g2	ANSPI	
g3	ARCIDIOCESI	
g4	ARTISTICO	
g5	ASSISTENZA	ENTE PATRONATO PREVIDENZA PUBBLICA
	ASSOCIAZIONE	GEOVA INVALIDI LUCRATIVA LUCRO MUSEO MUSICA MUSICALE NATURA ONLUS ORGANI PARTIGIANI PRO.LOCO PROMOZIONE.SOCIALE TEATRO TUTELA VEREIN VOLLEY
g6		VOLONTARIATO YOGA
g7	BASILICA	RETTORIA SANTUARIO
g8	STUDI	FONDAZIONE INIZIATIVE RICERCHE UNIVERSITA'
g9	CIRCOLO	ACLI AICS ALLEANZA ARCI ASSISTENZIALE AUSER CIRCOLO FENALC IPPICO LEGAMBIENTE NAUTICO RICREATIVO
g10	CLUB	ASD FOOTBALL LIONS ROTARY SPORTING TENNIS
g11	COMUNITA'	TERAPEUTICA
g12	CORPO	BANDISTICO VIGILI VOLONTARIO
g13	DIOCESI	OPERE SEMINARIO
g14	DIPENDENTI	CRAL DOPOLAVORO FONDO.PENSIONE
g15	EQUESTRE	ORDINE SCUDERIE TURISMO
g16	FITNESS	CENTER PALESTRA
g17	GENITORI	COMITATO INFANZIA LICEO RIABILITATIVO SCUOLA
g18	GRUPPO	AGESCI ANZIANI CINOFILO DANZE FOLKLORISTICO LEGA.NORD PALLAVOLO POPOLO TEATRALE
g19	ISTITUTO	FORMAZIONE ISTRUZIONE SOSTENTAMENTO
g20	LAVORATORI	AZIENDALE METALMECCANICI SINDACATO
g21	LIBERTAS	POLISPORTIVA
g22	LISTA	VALORI
g23	MISERICORDIA	CONFRATERNITA CONGREGAZIONE CONVENTO FRATERNITA'
g24	MUTUO.SOCCORSO	FREIWILLIGE OPERAIA
g25	OPERA.PIA	CASA.DI.RIPOSO COLLEGIO IPAB MONASTERO OSPEDALE
g26	PARROCCHIA	CANONICATO ENTE.ECCLESIASTICO PFARREI
g27	PARROCCHIALE	ABBAZIA BENEFICIO CHIESA PREBENDA
g28	PARTITO	COMUNISTA DEMOCRATICO ITALIANO SOCIALISTA
g29	PESCATORI	SPORTIVI
g30	POLIFONICO	CORO CURIA
g31	POLITICO	CONFERAZIONE MOVIMENTO
g32	PROTEZIONE.CIVILE	NUCLEO RADIO
g33	SANGUE	DONATORI FRATRES VOLONTARI
g34	SEZIONE	AVIS CACCIA COMBATTENTI DEMOCRAZIA FEDERCACCIA SINISTRA
g35	SINDACALE	RAPPRESENTANZA UNIONE
g36	SOCIETA'	ATLETICA BOCCIOFILA CASA.DI.CURA CICLISTICA COOPERATIVA NON.STATALE OSPEDALIERA PESCA
g37	UIL	CGIL CISL
g38	VIRTUS	BASKET PALLACANESTRO.

Fig. 3 Results using AP algorithm

The PAM solution creates still groups that are better than AP algorithm. The disadvantage is that, when the group is very large, it requires a new partition of cluster into subclasses. The centroid "associazione" (association) incorporates, e.g., several kinds of associations "associazione sportiva (sporting association), "associazione religiosa" (religious association), "circolo" (social club), etc. (Fig. 4).

The results show that the better answer to this issue is to find a solution by PAM method, using the number of groups selected by AP algorithm. Data processing has been provided by R software. We used the libraries Cluster (Maecheler et al. 2011), and Apcluster (Bodenhofer et al. 2011).

CLUSTER	CENTROIDS	KEYWORDS
g1	ABBAZIA	AMICIZIA ARCIDIOCESI BASILICA CANONICATO CARITAS ENTE ECCLESIASTICO GEOVA NON STATALE OPERE PARROCCHIA PARTIGIANI PATRONATO PFARREI POPOLO RETTORIA;
g2	ACCADEMIA	DIOCESI EQUESTRE ORDINE SCUDERIE TURISMO ;
g3	ACLI	AICS ARCI CIRCOLO FENALC IPPICO RICREATIVO;
g4	AGESCI	ANZIANI CINOFILO DANZE FOLKLORISTICO GRUPPO LEGA NORD ORGANI TEATRALE;
g5	ALLEANZA	MOVIMENTO POLITICO
g6	AMATORI	ATLETICA BASKET BILIARDO CALCIO PALLAVOLO RUGBY VOLLEY
g7	AMICI	MUSEO MUSICA NATURA PRO LOCO PROMOZIONE SOCIALE TEATRO
g8	ANSPI	GIOVANILE ORATORIO
g9	ARTIGIANATO	ARTISTICO PATTINAGGIO PERFEZIONAMENTO
g10	ASD	CENTER FITNESS PALESTRA
g11	ASSISTENZA	ENTE PUBBLICA
g12	ASSISTENZIALE	CRAL DIPENDENTI DOPOLAVORO FONDO PENSIONE PREVIDENZA
g13	ASSOCIAZIONE	CULTURALE DILETTANTISTICA
g14	AUSER	LEGAMBIENTE ONLUS VOLONTARIATO
g15	AVIS	CACCIA COMBATTENTI DEMOCRAZIA FEDERCACCIA SEZIONE SINISTRA
g16	AZIENDALE	LAVORATORI METALMECCANICI SINDACATO UNIONE
g17	BANDISTICO	CORPO MUSICALE VIGILI VOLONTARIO
g18	BENEFICIO	CHIESA PARROCCHIALE PREBENDA
g19	BOCCIOFILA	CICLISTICA COOPERATIVA PESCA SOCIETA
g20	CASA DI CURA	CASA DI RIPOSO COLLEGIO CONFERAZIONE FONDAZIONE FREIWILLIGE INVALIDI IPAB MONASTERO OPERA PIA OSPEDALE RIABILITATIVO SEMINARIO VEREIN
g21	INIZIATIVE	RICERCHE STUDI UNIVERSITA
g22	CGIL	CISL RAPPRESENTANZA SINDACALE UIL
g23	CLUB	DANCE FOOTBALL LIONS NAUTICO ROTARY SPORTING TENNIS
g24	COMITATO	GENITORI INFANZIA LICEO SCUOLA
g25	COMUNISTA	DEMOCRATICO ITALIANO PARTITO SOCIALISTA
g26	COMUNITA	
g27	CONFRATERNITA	CONGREGAZIONE CONVENTO FRATERNITA LUCRATIVA MISERICORDIA SANTUARIO
g28	CORO	POLIFONICO
g29	CURIA	FORMAZIONE ISTITUTO ISTRUZIONE OSPEDALIERA SOSTENTAMENTO TERAPEUTICA TUTELA YOGA
g30	DONATORI	FRATRES SANGUE
g31	LIBERTAS	PALLACANESTRO POLISPORTIVA VIRTUS
g32	LISTA	LUCRO VALORI
g33	MUTUO SOCCORSO	OPERAIA
g34	NUCLEO	PROTEZIONE CIVILE RADIO VOLONTARI
g35	PESCATORI	SPORTIVI

Fig. 4 Results using PAM algorithm

5 Conclusions

Text mining technique opens a new scenario for the treatment of textual documents registered in the administrative archives. The implemented procedure for the automatic classification of organizations name has worked effectively. Actually, knowledge engineering is labor intensive but, in the same time, it may ensure more precision and control over the classification process as a whole. However, it is relevant to empirically verify the advantages of different approaches. The results of the experimentation showed that, in the first time, it is relevant to classify automatically the corpus and to create a classes of expressions. This method helps us to detect more significant associations. The clustering produced with PAM and/or AP methods required a revision a posteriori. In the second time, we could apply Text categorization or Neural Network to classify NPI (Duda et al. 2000).

References

Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
Bodenhofer, U., Kothmeier, A., Hochreiter, S.: APCluster: an R package for affinity propagation clustering. Bioinformatics **27**, 2463–2464 (2011)
Bolasco, S.: Statistica testuale e text mining: alcuni paradigmi applicativi. Quaderni di Statistica **7**, 17–53 (2005)
Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications, 2nd edn. Springer, New York (2005)
Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2000)
Feldman, R., Sanger, J.: Text Mining Handbook. Cambridge University Press, Cambridge (2006)
Frey, B.J., Dueck, D.: Clustering by Passing Messages Between Data Points. Science **315**(5814), 972–976 (2007)
Iezzi, D.F.: Topic connections and clustering in text mining: an analysis of the JADT network. Stat. Anal. Textual Data Rome **2**(29), 719–730 (9–11 June 2010)
Iezzi, D.F.: Intimate femicide in Italy: a model to classify how killings happened. In: Palumbo, F., Lauro, C.N., Greenacre, M.J. (eds.) Data Analysis and Classification, p. 85–92. Springer, Berlin (2010). ISBN/ISSN: 978-3-642-03738-2. doi:10.1007/978-3-642-03739-9
Iezzi, D,F.: Centrality measures for text clustering. Commun. Stat. Theory Methods **41**, 3179–3197 (2012a)
Iezzi, D.F.: A new method for adapting the k-means algorithm to text mining. Ital. J. Appl. Stat.236 **22**(1), 69–80 (2012b)
Iezzi, D.F., Mastrangelo, M., Sarlo, S.: Text clustering based on centrality measures: an application on job advertisements. In: 11es Journées Internationales d'analyse statistique des données textuelles, pp. 515–524. Liegi, Belgium 13–15 giugno 2012
Istat: Istituzioni Nonprofit in Italia. Istituto nazionale di statistica, Roma (2001)
Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. Wiley, New York (1990)
Maecheler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: Cluster Analysis Basics and Extensions, R package version 1.14.1 (2011)
Rousseeeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987)
Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surveys **34**(1), 1–47 (2002)
Yang, T.: Computational verb decision trees. Int. J. Comput. Cogni. **4**(4), 34–46 (2006)

Linking Administrative Tax Records and Survey Expenditure Data at the Local Level

Lisa Crosato, Mauro Mussini, Paolo Mariani, and Biancamaria Zavanella

Abstract

The combined use of administrative data and sample surveys has become increasingly important in statistical applications. In this paper, we focus on the combination of administrative tax records and sample survey data collecting information on household expenditures. Our purpose is to create a new dataset containing information from both data sources following an almost-exact matching approach.

1 Introduction

Household surveys carry information about income, expenditures and several socio-demographic variables and are available at national and sub-national level. It is unlikely, though, that a survey designed for a specific project may be of use to a larger extent since it would probably miss some variables beyond the scope of the survey. In this case, suitable administrative sources can be used to cover the lack of information on the variables of interest. On the other hand, administrative agencies usually collect data that are relevant for administrative purposes. This implies that administrative data often do not correspond exactly with the survey variables and provide information which is not available in survey responses. When a research issue requires information available from different data sources (e.g., survey and

L. Crosato (✉) • P. Mariani • B. Zavanella
Department of Economics, Management and Statistics, University of Milan Bicocca, Milan, Italy
e-mail: lisa.crosato@unimib.it; paolo.mariani@unimib.it; biancamaria.zavanella@unimib.it

M. Mussini
Department of Economics, University of Verona, Verona, Italy
e-mail: mauro.mussini@univr.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_14,

administrative sources) the record matching of data from different sources can shift the analysis to a deeper level with respect to the original sources.

In this paper, we combine two data sources containing, among others, the main monetary proxies of well-being: income and consumption. Both sources, a sample survey and an administrative register, regard the municipality of Milan (Italy). The first dataset is the sample survey on family expenditure[1] (ICFM, hereafter) conducted by the Milan Municipality and the Chamber of Commerce of Milan (wave 2007–2008). The second dataset collects administrative records regarding the entire population resident in Milan and covers the period from 2000 to 2007. The administrative data derive from tax records matched to the local population and family register in the data warehouse AMeRIcA.[2] The integration with the population register represents one of the main advantages of AMeRIcA, since it allows the grouping of individuals in households and implies the total coverage of the population considering either individuals or households. Furthermore, several socio-demographic characteristics are recorded for tax-payers and not. This makes it possible also as a thorough characterization of citizens exposed to poverty, each year and over time (Crosato and Zavanella 2010; Minotti et al. 2010).

This study concerns the methodological aspects of the linkage between the above sample survey and the administrative register. We mainly focus on data combination, referring to the literature which offers several record linkage tools (Ceccarelli et al. 2008; Fortini et al. 2001; Gill 2001). Not disposing of a unique identifier for either households or individuals, we apply a relaxed version of the exact matching method (defined almost-exact matching by Gill (2001)) considering as linked two entities showing the same value over a given number of variables. We follow Jenkins et al. (2008) and combine the matching variables in four criteria, testing their discriminating power and reliability through several matching exercises. Even though the information stored in both datasets is confined to a single Italian municipality, our findings may have broader significance, especially on the methodological side, considering that all municipalities have access to the tax records of their residents and many of them conduct their own expenditure surveys. The paper is structured as follows. Section 2 describes the data sources. Section 3 introduces the linkage strategy by constructing and discussing matching variables (3.1) and finally presents linkage results (3.2). Section 4 concludes.

2 Data Description

We start by a brief but necessary description of the two data sources to be combined, since they are of a local character and so not very well known.

[1] ICFM stands for "Indagine sui Consumi delle Famiglie nel Comune di Milano".

[2] AMeRIcA ("Anagrafe Milanese e Redditi Individuali con Archivio") has been managed by the Statistics Department of Milan-Bicocca University on the behalf of Milan Municipality.

2.1 The ICFM Survey

The ICFM survey, referring to the 2007–2008 period, represents the survey data source that we use in this linkage exercise. This survey has been carried on with annual periodicity since 2005. The ICFM survey is based on a stratified sampling procedure using households as sampling units and the municipality sub-areas, the household size and the age of the householder as stratification variables. The questionnaire is very detailed, asking respondents to provide information on various forms of non-durable and durable consumption goods. In addition, the ICFM survey contains information on the properties lived in or owned by the household and the characteristics of each family member. The questionnaire also includes a question about the household net income level asking respondents to state their income class. Information in the 2007–2008 survey was collected from March 2007 till February 2008 and involved 808 households, including 2,403 individuals. A unique identifier was assigned to each household. However, it cannot be used for linkage purposes since it is a survey specific coding which is unrelated to any coding defined in other survey or administrative data (e.g. the unique identifier assigned to each household in AMeRIcA).

2.2 The AMeRIcA Data Warehouse

Administrative records derive from AMeRIcA project which provides demographic and income information concerning individuals and households resident in Milan. The structure of AMeRIcA is based on a data warehouse that combines administrative data records from the tax register of the Milan Revenue Agency with the Milan Population Register. Therefore, AMeRIcA is a combined administrative data source, and it represents the first experience in Italy of linkage of administrative records belonging to different administrative data sources. The Revenue Agency registers the Italian Personal Income Tax (IRPEF) in categories: pre-tax income by source (pre-tax income is defined as the sum of incomes from dependent employment, self-employment, company income, rental income, agricultural income, real estate income and other incomes), taxable income, gross IRPEF, net IRPEF, tax allowance, detraction for dependent family members and other deductible costs. The archives of the Registry provide information about marital status, citizenship, address, family composition, residence and other personal data (gender, date of birth) for each registered individual. The combination of the two administrative sources enables AMeRIcA to render information on pre-tax income, income tax, tax allowance and detractions plus family size and composition of the residents in Milan. The AMeRIcA pilot project began in 2000 on the basis of the rising interest for the use of administrative data for statistical purposes, and now the data warehouse covers the 2000–2007 period. In our study, we focus on household records available for 2007. We refer to a population that is composed of 653,686 households.

3 Data Linkage

Statistical literature offers three main tools for record linkage: statistical matching, probabilistic matching and exact (or deterministic) matching (see Gill 2001). Statistical matching is advisable if the fraction of units that are in both data sources is small, so that we can treat the two samples as independent with negligible intersection. In our case exact or probabilistic matching are the most suitable methodologies, since not only a large part of the units are common to both data sources, but households belonging to the survey are sampled from the administrative source. Therefore the first dataset is, on principle, included in the second. A possible source of violation of the above statement could be mainly information recorded at different times, such as births or deaths after the interview for the survey has been made or residential mobility acknowledged only by the population register. Therefore, we assume in the following all units in the ICFM survey to be included in AMeRIcA. In record linkage literature, this situation is named nested linkage and it has been studied by Copas and Hilton (1990), who fully discussed the statistical problems concerning record linkage, and proposed a probabilistic approach using a training set of known correctly linked records to estimate likelihood ratios for matching.

Differently from Copas and Hilton, we face nested linkage by adopting an exact matching approach. Exact matching is feasible when both datasets contain the same variable or characteristic available for all units, fixed, easily recordable, verifiable and unique to that individual (Gill 2001). When available, this variable is usually identified with some unique identification number assigned to individuals at birth such as the National Health Service Number or, in Italy, the fiscal code, an alphanumerical code composed by letters and numbers corresponding to name, sex, date and place of birth. An important precedent of data using exact matching with fiscal code to combine income and consumption data was conducted by Ceccarelli et al. (2008). The above authors matched several Eu-Silc waves with the Italian tax register in order to assess measurement error in Eu-Silc regarding Italy and to retrieve missing information from the tax register. They had the opportunity to access micro data on tax records at the national level and to assign fiscal codes to individuals present in both datasets. On the contrary, for privacy reasons, we do not possess either the fiscal code or the information useful to retrieve it in either data source. Therefore, in our case exact matching would not be feasible due to the absence of a unique identifier.

When the data sources lack a unique identification variable, the alternative is to artificially create a compound key by starting from a few characteristics which jointly form a sort of identifying code of each unit (Gill 2001). In other words, the records under scrutiny are compared simultaneously on several variables, so their different combinations constitute many criteria to proceed with (Jenkins et al. 2008). Due to this highly characterizing matching key, exact matching should in principle return one-or-none type links, leaving no space for the so called possible links. Since we could not identify a set of variables that jointly substitute the unique

identifier, we resort to almost-exact matching as defined by Gill (2001) by relaxing the exact match criterion. We use the number of variables that agree (at least three) to establish if a record pair should be linked. In practice, we follow Jenkins et al. (2008) procedure also to have a basis for comparison. However, it should be stressed that with respect to the above authors, we maintain the possible link category among our outcomes.[3]

3.1 Variable Selection and Matching Criteria Definition

Before proceeding further with the linkage, it is worthwhile to devote a few lines to the choice of matching variables and matching units in order to clarify the pattern we followed.

The variables both (potentially) common and unique to the ICFM survey and the data warehouse AMeRIcA are reported in Table 1. The first kind of variables represents candidate matching variables, where candidate refers to the way the same variables are defined and/or recorded. Among these, we left apart those related to income since they originate from completely different processes: income and the number of income receivers reported in the survey derive from specific questions whereas in AMeRIcA the same variables derive from tax records. These and the second kind of variables (i.e. those which are not common to both datasets) represent the informative added value in which either dataset brings through the matching process: the first ones for comparison and possible correction of item non-response; the second ones for enlarging the informative set characterizing households and individuals. Other variables potentially useful in data matching, but differently recorded, are the characteristics of households, which in the survey are collected or double-checked during the interview and in AMeRIcA are registered through administrative procedures.

Table 1 is also divided into two parts according to the candidate matching units, since either dataset supplies a number of characteristics for both households and individuals. As a consequence, our first approach was to separately process individuals and households matching, proceeding to the linkage within individuals (or households) on the basis of their proper characteristics, but this unfortunately resulted in not a single positive link or negative link, only possible links. Starting from households, we tried a matching using Enumeration Area, Number of Components and Number of Children, after blocking according to the Type of Family. Of course, the Number of Children was of no help in identifying families with no children and singles, but in any case, also for large (with 6 or 7 components) families, results were very poor. It went even worse with our attempt to link individuals according to Sex, Age and Enumeration Area retrieved from the family. Then, in order to augment the quality of our matching variables, we decided to fully

[3] In this linkage exercise, possible link occurs when one record in the ICFM dataset is linked to many records in the AMeRIcA dataset; namely, there exists a one-to-many type of link.

Table 1 Comparison of variables included in AMeRIcA and in the ICFM survey

ICFM survey	AMeRIcA data warehouse
Matching units: households	
Postal code (CAP: 38 codes)	Address
Enumeration area	Functional area (180 areas)
	Enumeration area (6,036 sections)
Number of components	Number of components
Number of children	Number of children
Type of family (1 = single, 2 = couple with children, 3 = couple without children, 4 = single parent)	Type of family (1 = single, 2 = couple with children, 3 = couple without children, 4 = single parent, 5 = other)
Number of income receivers	Number of income receivers
Professional condition of the householder	
Number of pension recipients	
Monthly consumption	
Income class	Taxable income
	Net income
	Taxes
	Pensions and subsides
	Number of Italians
	Income source
Matching units: individuals	
Sex	Sex
Year of birth (age as 2007-year of birth)	Age
Reference person (Householder)	Householder
Relationship of each family component to the householder	
Education level	
Working position	
	Taxable income
	Net income
	Taxes
	Pensions and subsides
	Income source

Source: ICFM survey and AMeRIcA data warehouse

exploit our data sources and in particular to integrate matching units constructing two artificial variables relating households and individuals. Each household was assigned two vectors:

- An Age Code reporting ages of all family components in increasing order;
- A Sex Code given by the sex of all family components ranked according to increasing age.

Of course, these artificial variables share potential reporting errors in the number of components, and the family sex code is affected by errors in age recording when implying reverting components' age ranking. Apparently, the Age Code and

Sex Code possess a large discriminating power but may present some problems of reliability, since they can suffer from an accumulation of errors.

The other two variables available on households are Type of Family and Enumeration Area. The first one might cause some mismatch due to the category "other", which is present in AMeRIcA but not in the ICFM survey. On the contrary, Enumeration Area should be measured in the same way in both datasets (being assigned and not requested in the survey) and subject, at most, to coding or reporting errors and to residential mobility discrepancies. In addition, Enumeration Areas represent the finest territorial grid of Milan municipality with 6,036 sections. For these reasons we expect it to show both high reliability and discriminating power. To conclude, we single out four variables suitable for our matching exercise on households and precisely Enumeration Area, Type of family, Age Code and Sex Code. Following Jenkins et al. (2008), we organize the selected variables in four criteria, each of them excluding one variable at a time. The main advantage of this procedure is that, as we do not dispose of a unique identifying code, using the single variables rarely leads to match any record, while using compound keys (or criteria) allows for some success in matching and at the same time the exclusion of one variable in turn allows for assessment of their discriminating power.

The four linkage criteria are:

- Criterion 1 (C1): Enumeration Area, Age Code and Type of Family;
- Criterion 2 (C2): Enumeration Area, Sex Code and Type of Family;
- Criterion 3 (C3): Type of Family, Sex Code and Age Code;
- Criterion 4 (C4): Enumeration Area, Sex Code and Age Code.

We first run the four linkage exercises independently and secondly combine the criteria in a hierarchical matching process (HM, hereafter). By applying HM separately for possible matches and non-matches, one can isolate the capability of the second criterion to solve uncertainties (possible matches given by the first criterion) and to correct non-matches identified by the first criterion. The HM may be useful to evaluate the consistency of the various matching criteria in determining matches, possible matches and non-matches. Moreover, it provides preliminary indications for manual revision since it contributes to isolating misclassified record pairs. For instance, a record pair that is stated as match by a criterion but not by another can be checked by manual review in order to establish its true status.

3.2 Linkage Rates

Table 2 reports linkage results both in absolute values and as a fraction of number of records in the ICFM survey. Table 2 is divided into three main panels. The top one refers to linkage according to single criteria and to the pooled linkage obtained using at least one of the criteria. The central panel reports linkage results obtained by applying HM on possible matches. The bottom panel reports further matches resulting from application of HM on non-matches. Among the four independent criteria, C1 (composed of Enumeration Area, Age Code and Type of Family) and C4 (composed of Enumeration Area, Sex Code and Age Code) return linkage rates

Table 2 Linkage rates for ICFM households

	Matches		Possible matches		Non-matches		All
Independent matching	*n*	%	*n* (to *N*)	%	*n*	%	*n*
C1	509	63.0	91 (*254*)	11.3	208	25.7	808
C2	138	17.1	584 (*9,299*)	72.3	86	10.6	808
C3	204	25.3	476 (*338,624*)	58.9	128	15.8	808
C4	562	69.6	57 (*143*)	7.1	189	23.3	808
Pooled matching	615	76.1	149	18.4	44	5.5	808
HM on possible matches							
C1 + C2 or C1 + C4	541	67.0	57 (*143*)	7.1	210	26.0	808
C1 + C3	526	65.1	73 (*307*)	9.0	209	25.9	808
C2 + C1 or C2 + C4	535	66.2	67 (*166*)	8.3	206	25.5	808
C2 + C3	390	48.3	247 (*6,049*)	30.6	171	21.2	808
C3 + C1	468	57.9	124 (*494*)	15.3	216	26.7	808
C3 + C2	369	45.7	259 (*6,149*)	32.1	180	22.3	808
C3 + C4	473	58.5	114 (*428*)	14.1	221	27.4	808
HM on non-matches[a]							
C1 + C2	567	70.2	169 (*1,230*)	20.9	72	8.9	808
C1 + C3	537	66.5	159 (*23,105*)	19.7	112	13.9	808
C1 + C4	584	72.3	57 (*143*)	7.1	167	20.7	808
C2 + C1	551	68.2	67 (*166*)	8.3	190	23.5	808
C2 + C3	393	48.6	252 (*6,070*)	31.2	163	20.2	808
C2 + C4	553	68.4	67 (*166*)	8.3	188	23.3	808
C3 + C1	485	60.0	124 (*494*)	15.3	199	24.6	808
C3 + C2	387	47.9	291 (*6,314*)	36.0	130	16.1	808
C3 + C4	498	61.6	114 (*428*)	14.1	196	24.3	808
C4 + C1	584	72.3	59 (*147*)	7.3	165	20.4	808
C4 + C2	585	72.4	155 (*1,272*)	19.2	68	8.4	808
C4 + C3	571	70.7	134 (*37,233*)	16.6	103	12.7	808

Notes: Linkage rates are calculated as a proportion of the ICFM survey ($n = 808$). The column entitled "possible matches" reports in brackets the number of records in AMeRIcA (N) possibly matching the n records in the ICFM survey. In HM, criteria are applied in order of appearance
[a]Results include the matches gained through HM on possible matches

remarkably larger than the remaining two. In particular, the best results are provided by C4 with a linkage rate of 69.6 % while C1 matches 63 % of the records, both expressed as a fraction of the ICFM survey. On the one hand, this suggests that Enumeration Area and Age Code are the outstanding variables in our linkage exercise. Comparing the column of possible linkage rates of C2 and C3 confirms the noticeable discriminating power of Enumeration Area versus Age Code since the lack of the first leads to the largest ratio between the number of records in AMeRIcA to be possibly matched to ICFM survey records and the number of the corresponding ICFM records (711 to 1 for C3 against almost 16 to 1 for C2).

The central panel of Table 2 reports the results of HM on possible matches (for instance, C1 + C2 means that C2 was applied to possible matches established by C1). We observe that rates corresponding to application of any criterion after C4 are not reported in Table 2 because there was no further match to be counted. On the contrary, we can see that in the remaining cases applying a second criterion solves many uncertainties, leading to a gain in terms of matching rate especially for C2 and C3. Overall, we can state that the contribution of HM greatly varies according to the goodness of the first criterion applied. Unsurprisingly, the worst rates are obtained from combination of C2 and C3 independently of their application order, confirming their low discriminating power. The case of HM applied to non-matches (Table 2, bottom panel) points to slightly different conclusions, since all linking rates corresponding to the four independent criteria are improved. Matching rates are to be compared with the central panel ones since they are obtained by counting all matches achieved both through the single criteria and through HM on possible links. Three combinations outperform the others, precisely C1 + C4, C4 + C1 and C4 + C2. All of them exceed 72 % of record matches, but mixing C1 and C4 seems the best solution, since it minimizes the possible match ratio to 2.5 records in AMeRIcA for 1 record in the ICFM survey. The return to HM with respect to C4 alone consists in 2.7 % of additional matches gained from the non-matches according to C4 (the number of possible matches in fact remains stuck to 57). Thus, it seems worth checking the differences in the left-out variable (Age Code or Type of Family) between additional matched records by manual revision.

To conclude, we observe that linkage rate varies between subsamples defined by the type of family. The lowest rates are achieved for singles and couples without children. Since our almost-exact linkage mainly relies on demographic variables concerning family members, the lower the household size the lower the discriminating power of the variables Age Code and Sex Code. This may introduce selection bias which should be taken into account when analysing linked data.

4 Conclusion

In this paper we have combined two different data sources: the ICFM survey conducted by the Milan Municipality and the Chamber of Commerce of Milan and the tax register matched to the local population and family register in the data warehouse AMeRIcA. Matched households are now endowed with information collected in both datasets. However, when using linked data, attention should be paid to the presence of selection bias introduced by the linkage procedure.

Furthermore, we have discussed and tested alternative variables to perform data combination, even in the absence of unique identifiers or highly identifying characteristics. From a methodological standpoint, this may have broader validity since the variables used in our linkage procedure (age, sex, type of family) are usually collected in expenditure surveys. Therefore, our linkage exercise could be replicated in other municipalities.

Future research could be devoted to compare the results of this study with those provided by probabilistic linkage (Copas and Hilton 1990) in order to review our preliminary results. Furthermore, linked data may offer new insights on poverty estimation and well-being indices construction at the local level.

Acknowledgement We thank Milan Municipality (Statistics Office) and the Chamber of Commerce of Milan for providing us with the data. We are particularly grateful to Lorena Scarcello and Flavio Necchi for their support to our project.

References

Ceccarelli, C., Coppola, L., Cutillo, A., Di Laurea, D.: Combining survey and administrative data in the Italian EU-SILC experience: positive and critical aspects. Retrieved in July 2010. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1683449 (2008)

Copas, J.B., Hilton, F.J.: Statistical models for matching computer records. J. Roy. Stat. Soc. Ser. A **153**, 287–300 (1990)

Crosato, L., Zavanella, B.: L'evoluzione del reddito dei cittadini milanesi (2000–2004) sulla base di archivi amministrativi. In: Mezzanzanica, M., Zavanella, B. (eds.) I numeri della città: un quadro socio-economico del comune di Milano sulla base di fonti amministrative, pp. 125–180. FrancoAngeli (2010)

Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M.: On Bayesian record linkage. Res. Official Stat. **4**, 185–198 (2001)

Gill, L.: Methods for automatic record matching and linking and their use in National Statistics. National Statistics Methodological Series No. 25. Office for National Statistics, UK (2001)

Jenkins, S.P., Lynn, P., Jäckle, A., Sala, E.: The feasibility of linking household survey and administrative record data: new evidence for Britain. Int. J. Soc. Res. Meth. **11**, 29–43 (2008)

Minotti, S.C., Mussini, M., Zavanella, B.: Simulazione di alcuni sistemi fiscali europei sui redditi delle famiglie milanesi. In: Mezzanzanica, M., Zavanella, B. (eds.) I numeri della città: un quadro socio-economico del comune di Milano sulla base di fonti amministrative, pp. 181–221. FrancoAngeli (2010)

An Application of Statistical Matching Techniques to Produce a New Microeconomic Dataset on Farming Households' Institutional Sector in Italy

Edoardo Pizzoli, Benedetto Rocchi, and Giuseppe Sacco

Abstract

A new microeconomic database on farm households in Italy was created using statistical matching techniques. Information on total households' income and well-being gathered by the EU-SILC survey on living conditions for Italy was attached to the observations included in the FBS database for Italy. The new dataset, still representative of agriculture as an industry, also allows a proper statistical representation and socio-economic characterization of farm households as an institutional sector.

The quality of the new microeconomic information was assessed analysing the statistical properties of key analysis variables and the distributive features of the current UE Common Agricultural Policy.

1 Introduction

In carrying out insightful analyses of distributive implications of alternative agricultural policy options, suitable microeconomic information on potential beneficiaries is needed. Two main features seem to be relevant to the analysis. First, the institutional sector of farm households needs to be properly placed within the economy-wide income distribution, observing the total household income (THI) (Unece et al. 2007); second, information should be available to classify households both using information on the farm (such as size, product typology, management

E. Pizzoli (✉) • G. Sacco
ISTAT, Rome, Italy
e-mail: pizzoli@istat.it; sacco@istat.it

B. Rocchi
University of Florence, Firenze, Italy
e-mail: benedetto.rocchi@unifi.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_15,

form) and information on well-being of the household itself (such as composition, age, education, health).

The main sources of microeconomic information on the institutional sector of farm households, such as the Farm Business Survey (FBS) carried out by ISTAT or the European Farm Accountancy Data Network (FADN), fail to comply with both these characteristics: their focus on technical aspects and the centrality given to income from farming makes these surveys suitable for analysis only within an industry (agricultural) perspective.

This paper aims to propose a possible solution to this information problem. In the next paragraph a description of data and methods used in the analysis will be proposed. The assessment of new dataset produced through matching techniques with some exemplificative results will follow. In the subsequent paragraph a statistical analysis will be performed on key objective variables included in the datasets result of several tests. A final paragraph will show some figures on distributive features of the farming households' sector in Italy resulting from the new selected dataset.

2 Data and Methods

A new microeconomic database on farm households in Italy was created using statistical matching techniques (Rassler 2002; D'Orazio et al. 2006). Information on total households' income and well-being gathered by the EU-SILC survey on living condition for Italy (ISTAT 2010) was attached to the observations included in the FBS database for Italy (ISTAT 2011). The new dataset, still representative of agriculture as an industry, also allows a proper statistical representation and socio-economic characterization of farm households as an institutional sector (Rocchi 2010).

The FBS, designed to supply information for national accounts, yearly surveys a sample of agricultural holdings representative of the Italian agriculture. The database includes a detailed set of variables on farm structures (such as cultivated area, livestock number, labour employment) and on costs and revenues from farming. According to these information a good estimate of income from farming can be obtained. Furthermore, for the farm households (the largest part of the sample) a small set of variables on household's composition as well as on extra-farm source of income (classes of income by four types of sources) is available (Pizzoli 2005).

The EU-SILC is a sample of Italian households designed to gather detailed information on incomes as well as on living condition and well being. The sample is representative of total Italian population but, given the optimization criteria adopted in the design of the survey, *farm* households are under-represented (520 observations from a total of 20,982, that is 2.48 %). The dataset includes variables on occupation, professional position and income sources by type of single household's members; a number of nominal variables expressing well-being of household's members and family living condition are available as well.

Table 1 Matching variables

Variable name	Description	Continuous	Type
ncomp	Number of household members	No	Scale
ex_i	Extra-farm net income from self-employed labour	Yes	Scale
ex_d	Extra-farm net income from hired labour	Yes	Scale
ex_p	Extra-farm net income from pensions	Yes	Scale
ex_c	Extra-farm net income from capital assets	Yes	Scale
yagrn	Net income from farming	Yes	Scale
redtotale2	Total household income (THI)	Yes	Scale
Quex_i	Share of extra-farm income from self-employed labour	Yes	Scale
Quex_d	Share of extra-farm income from hired labour	Yes	Scale
Quex_p	Share of extra-farm income from pensions	Yes	Scale
Quex_c	Share of extra-farm income from capital assets	Yes	Scale
agrn	Net income from farming more than 50 % of THI	No	Binary
redtotale2_pc	Per capita total household income	Yes	Scale
decile	Income decile	No	Ordinal

Farm households in FBS and EU-SILC samples can be assumed to be homogeneous, as they represent the same typology of statistical units, and coming from the same target population, according to the following units definition: "households ... that derived any income, however minor, from agriculture or contributed some labour input to agricultural production". ("broad" definition, Chapter IX, The Agricultural Household—Concepts and Definitions; Unece et al. 2007). Farm households, by construction, belong to the larger households population and are a specific typology (socio-professional) group.[1]

For the aim of the analysis a sub-sample of 9,858 observations representative of 1,586,193 farm households from the FBS (year 2007) was considered as the "recipient" database. A set of 14 "matching variables" on households' characteristics was defined according to available information. The criterion followed in the variables selection was the possibility to exactly replicate them for each observation included in the "donor" EU-SILC database (year 2007). Table 1 lists the matching variables with some information on them.

[1] In the EU-SILC survey a "'private household' is a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of the essentials of living" (Art. 2, Definitions; EU 2003). This definition is equivalent to the UN definition (UN 1998) adopted by Eurostat and EU members countries.

A farm household in the FBS sample is defined as a household with at least a spouse that manages an unincorporated or quasi-corporate agricultural holding (individual farms; communal tenures), and works in the agricultural holding. A farm household in the SILC sample is defined a household with at least a family member earning incomes from self-employed labour in agriculture, according with individual records where incomes are classified by sector of economic activity. For a discussion on the definition of agricultural household see Chapter IX of the UN Handbook (UN 2011).

Table 2 Regional stratification of observations in the original datasets

Region	Frequency in recipient	Frequency in donor	Donor to recipient ratio
1	1,552	4,973	3.20
2	2,607	4,990	1.91
3	1,951	4,950	2.54
4	2,876	4,400	1.53
5	872	1,669	1.91
Total	9,858	20,982	2.13

Both donor and recipient samples were stratified according to a space variable (the region each to which observation belongs). Two different regional stratifications were assessed (5 and 20 regions corresponding to Nuts1 and Nuts2 classifications). To ensure a well-balanced stratification both in the recipient and in the donor database the 5 regions stratification was finally adopted. The result of layering is shown in Table 2.

The donor to recipient ratio shows a good distribution of the 20,982 donors with respect to the 9,858 recipients.

The integrated archive was built by means of statistical matching techniques based on nonparametric imputation methods (hot-deck). More precisely in the realization of the matching between the two files was used the method of nearest-neighbour imputation where the proximity between two records is expressed by an appropriate distance function.

The distance function chosen for the matching procedure is the mixed distance (Gower distance), in order to take into account the presence of discrete variables between the matching variables. Given the value assumed for the observations a and b by k variables x_j available in both databases,

$$\text{Gower} : \frac{1}{k}\sum_{j=1}^{k} c_j d_j\,(a,b)$$

where: for categorical variables: $c_j = 1$, $d_j(a,b) = 0$ if $x_{a_j} = x_{b_j}$ and 1 otherwise; for continuous variables: $c_j = 1/\text{Range}(x_j)$, $d_j\,(a,b) = \left|x_{a_j} - x_{b_j}\right|$

For each matching variable, the range is calculated considering the observations of both samples. Indicated with A and B respectively the set of possible values that can assume the variable x_j in the set of donors and in the set of recipients will have:

$\text{Range}\left(x_j\right) = x_{j_1} - x_{j_2}$, where $x_{j_1} = \max\left(x_{j_i}/x_{j_i} \in A \cup B\right)$ and $x_{j_2} = \min\left(x_{j_i}/x_{j_i} \in A \cup B\right)$

The matching was achieved by placing the constraint that a record could not be donated more than two or three times; have also been considered as donors not only those with minimum distance but all those who had a distance $d(a, b)$ within the range:

Table 3 Parameters adopted in the replications of matching

Name	Maximum number of donations	w_{thi}
Test1	2	0.50
Test2	3	0.50
Test3	2	0.75
Test4	3	0.75

$$d_{\min} - 0.01 \leq d\,(a, b) \leq d_{\min} + 0.01$$

where $d(a, b)$ is the observed distance between the units a and b and $d_{\min}$ is the minimum distance observed.

Different weights can be assigned to the matching variables. Given the aim of the analysis (to create an improved dataset to ground the estimate of the total income of farm households) in the matching procedure the largest weight was assigned to *redtotale2*, the variable representing the THI, respectively 0.50 or 0.75.[2] Given w_{thi} the weight assigned to THI, a weight equal to $(1 - w_{\text{thi}})$ was equally subdivided among the other matching variables.[3]

Combining the maximum number of donation of the same record from donor dataset with the two set of weights results in 4 replications of the matching procedure according to Table 3.

3 Statistical Checking

To reliable final analysis of results from statistical matching procedure, it is important to check if small changes of the key matching variable (THI) weight, w_{thi}. significantly affect the parameters of the resulting distribution for the variables of analysis generating probable unstable results. This statistical checking has been carried out on *ncomp*, the control variable available for all the datasets (including FBS), and *redtotale2_pc*, the objective variables of the Tests.

Descriptive statistics for *ncomp* and *redtotale2_pc* are reported in Table 4.

The distributions of the two variables have the same shape, asymmetric with long tails to the right-end-side and more peaked than normal distribution. Parameters slightly change with the different Tests.

[2]So far the choice of the distribution weight was oriented by *a priori* considerations of the subjective nature: it was thought that the variable THI (redtotale2) should be more decisive in the matching process. In a future development the correlation between matching variables and the variables of interest in the donor database will be assessed as a possible criterion in choosing distribution weights.

[3]The software package used in this paper was originally built for the production of an integrated archive for the social accounting matrix of Italian economy. A short documentation on the software is available in the Manual (Sacco 2008) at the site http://cenex-isad.istat.it.

Table 4 Descriptive statistics (valid $N = 9858$)

Var.	Mean	Confid. −95.0 %	Confid. +95.0 %	Median	Min.	Max.	SD	SK	KU
ncomp									
FBS	1.7736	1.7552	1.7921	2	1	16	0.93548	1.859	9.304
Test1	1.5021	1.4875	1.5167	1	1	6	0.73799	1.537	2.395
Test2	1.5346	1.5199	1.5494	1	1	6	0.74648	1.419	1.895
Test3	1.5371	1.5223	1.5518	1	1	6	0.74704	1.421	1.969
Test4	1.5469	1.5321	1.5618	1	1	6	0.75257	1.414	1.937
redtotale2_pc									
Test1	16,146.4	15,921.3	16,371.4	13,594.8	0	86,165.4	11,398.9	0.876	0.302
Test2	16,595.4	16,373.6	16,817.2	13,594.8	0	92,923.2	11,236.5	0.926	0.852
Test3	16,609.7	16,386.9	16,832.5	13,594.8	0	86,866.1	11,284.1	0.957	1.045
Test4	16,639.1	16,415.4	16,862.9	13,672.5	0	94,167.2	11,334.0	1.018	1.445

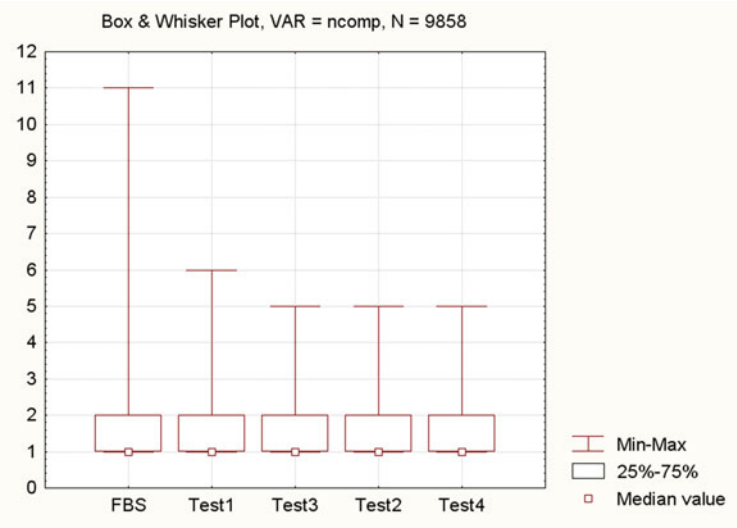

Fig. 1 Box and Whisker plot for the number of household members

Considering the number of household members, FBS variable has a much higher range and variability, while Test2, Test3 and Test4 variables are very close (see Fig. 1)

Considering per capita THI, all the Tests variables are very close and only Test4 variable shows a higher range of variability (see Fig. 2).

A *t*-test has been used to evaluate the differences in means between pairs of variables (Table 5).

Considering the first variable, *ncomp*, the *p*-levels reported suggest that the research hypothesis about the existence of a difference in means can be accepted,

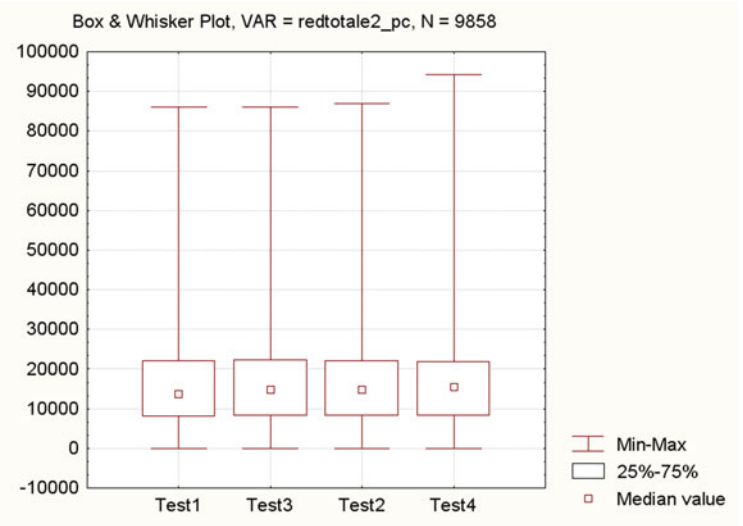

Fig. 2 Box and Whisker plot for per capita total household income

Table 5 *t*-Test for difference in means, independent samples[a] (valid $N = 9858$)

Var.	*t*-value	*df*	*p*	*F*-ratio variances	*P* variances
ncomp					
FBS vs.Test1	22.62803705	19,714	0.000000000	1.606811236	0.00000000
FBS vs.Test2	19.82692898	19,714	0.000000000	1.570483757	0.00000000
FBS vs.Test3	19.61920737	19,714	0.000000000	1.568120907	0.00000000
FBS vs.Test4	18.74891138	19,714	0.000000000	1.545132341	0.00000000
Test1 vs. Test2	−3.079968607	19,714	0.002073072	1.023131394	0.25631243
Test1 vs. Test3	−3.308986123	19,714	0.000938023	1.024673052	0.22632157
Test1 vs. Test4	−4.223467879	19,714	2.41651E-05	1.039918195	0.05202248
Test2 vs. Test3	−0.228886566	19,714	0.818959462	1.001506804	0.94042060
Test2 vs. Test4	−1.149698167	19,714	0.25028213	1.016407278	0.41918106
Test3 vs. Test4	−0.921314689	19,714	0.356897437	1.014878056	0.46349586
redtotale2_pc					
Test1 vs. Test2	−2.78528	19,714	0.005353	1.029113	0.154293
Test1 vs. Test3	−2.86820	19,714	0.004133	1.020455	0.314840
Test1 vs. Test4	−3.04351	19,714	0.002341	1.011479	0.570993
Test2 vs. Test3	−0.08936	19,714	0.928797	1.008485	0.674905
Test2 vs. Test4	−0.27205	19,714	0.785583	1.017434	0.390916
Test3 vs. Test4	−0.18251	19,714	0.855185	1.008873	0.660998

[a]The selected samples, from FBS survey and the four tests, are assumed to be independently generated with respect to the two objectives variables: number of household members (*ncomp*) and per capita total household income (*redtotale2_pc*). Relaxing this assumption, the power of *t*-test should be considered with respect to other tests

Table 6 Extra-farm income estimates (Mio€, 2007)

	FBS	Test1	Test2	Test3	Test4
ex_i	3,005	6,038	6,180	6,401	6,197
ex_d	6,956	12,898	13,477	13,547	13,688
ex_p	6,801	4,077	4,345	4,375	4,359
ex_c	151	4,831	5,358	5,378	5,356
Total	16,914	27,844	29,360	29,702	29,600

as expected, comparing FBS variable with the Tests variable. Test1 variable also do not pass the test with respect to the other Tests variable.

Considering the second variable, *redtotale2_pc*, the same results are confirmed between the Tests variable.

If a one-way ANOVA is computed on all four *ncomp* and *redtotale2_pc* Tests variables, the previous results are confirmed only for the first variable: *ncomp* mean in Test1 significantly differs from the other means in Test2–4 (F-value = 6.71, $p = 0.0002$), while for *redtotale2_pc* the means can be considered not significantly different (F-value = 2.48, $p = 0.0589$). If the same analysis is replicated only on Test1–3 variables, the null hypothesis of equal mean can be accepted at 5 % significance level for both *ncomp* and *redtotale2_pc*.

This is an indicative result for the matching procedure: Test1 considers a weight for the key matching variable equal to 0.5, and changing the donation from 2 (Test2) to 3 (Test1) can change the mean estimation. A higher weight for THI assures a greater stability of results.

Finally, a comparison among the four final databases is proposed in Table 6. Each row shows alternative estimates of the total extra-farm income by different sources. The first column displays the totals that could be estimated using only information included in the original FBS database[4] while the others show the estimates obtained using information originally included in the EU-SILC database and "matched" with FBS records according with the procedure described above.

Two relevant results can be stressed. First, the use of the new database, whatever the replication considered, leads to a quite different estimate of totals (namely, larger for self-employed labour, hired labour and capital asset incomes, and smaller for income from pensions); second, the outcome of the matching procedure does not seem to be sensibly affected by changes in the parameters (max number of donations and weights assigned to matching variables).

[4]More precisely the estimate was based on the matching variables. In the FBS only classes of extra-farm incomes (by source) are collected. To estimate the absolute value of each income component an average value was associated to each class. The average value of each class for each income component was estimated through regression using the EU-SILC database, where single income sources are collected in absolute value, and used to prepare the matching variables both in the recipient and in the donor database.

Table 7 Comparison among alternative matching results

	Test1	Test2	Test3	Test4
Total income: combined vs. matched				
Percentage difference	45.2	42.7	42.3	42.4
Correlation	0.585	0.596	0.626	0.610
Average Gower distance	0.043	0.044	0.036	0.046

A further comparison between the four replications is proposed in Table 7. In the first row the total income of households estimated combining the income from farming from FBS data with extra-farm income matched from EU-SILC, is compared with the THI of "donor" observations (as quantified in the original EU-SILC dataset). Not surprisingly the large, positive percentage difference shows that the farming component of total income would be underestimated using EU-SILC data alone. Test1 shows a larger difference (>45 %) in the estimate of totals, while the other three replications lead to quite similar results. The best correlation between "combined" and "matched" total income variables is shown by Test3. Finally, the average value of the Gower distance in the space of the matching variables between recipient (FBS) and donor (EU-SILC) records is proposed in the last row. Again, the best performance is shown by Test3, the only one with an average distance lower than 0.04.

4 Some Preliminary Results

Overall, these results seem to show that a relevant information may be added to the original FBS using statistical matching techniques. Furthermore, the matching procedure yields results quite robust in front of variation in the values of parameters. To highlight the potential interest of the matching experiment in this paragraph the Test3 database is used to estimate some figures on the distributive features of the farming households' sector in Italy.

In Table 8 some figures on the distributive features of the farming households' sector in Italy are displayed. Families are classified according to the prevalence of income from farming[5] (agricultural vs. non-agricultural) and by income quintile. The reader should bear in mind that income quintiles were defined taking into account the whole Italian population, not only the sector of farm households. As a consequence in Table 8 the households are not equally distributed among quintiles: figures in the first column show the position of households managing agricultural activities in Italy within the *overall* income distribution.

For the largest part of households involved in agriculture farming is only a secondary source of income. "Agricultural" households in a narrow sense (income from farming is more than 50 % of THI) are less than 20 %. Noticeably, in the

[5] A household is classified as "agricultural" if farming supplies more than 50 % of the THI.

Table 8 Distributive features of the farming households' sector in Italy, 2007

Income quintile	Percentage of households	Average per capita income (€)	Percentage of net income from farming	Percentage of SFP	SFP/total household income (%)	SFP/net income from farming (%)	Well-being index
Non-agricultural 1	43.1	7,714	9.3	11.4	2.9	26.5	5.9
Non-agricultural 2	15.2	14,777	6.2	6.4	3.0	22.3	6.4
Non-agricultural 3	12.1	19,218	6.6	6.1	3.2	19.9	6.7
Non-agricultural 4	6.8	24,713	6.9	6.5	4.6	20.3	7.5
Non-agricultural 5	5.7	46,925	10.4	11.0	5.1	23.0	8.5
Agricultural 1	7.7	4,557	5.4	6.3	18.9	25.1	6.0
Agricultural 2	1.5	14,619	3.2	2.7	11.7	18.7	6.3
Agricultural 3	1.3	19,252	3.6	3.5	13.2	20.8	6.7
Agricultural 4	1.3	25,523	4.4	4.2	12.6	20.6	7.4
Agricultural 5	5.3	66,266	44.0	41.8	14.9	20.5	8.3
Total	100.0	16,904	100.0	100.0	6.4	21.6	6.5
Q5/Q1 agr	0.1	6.1	1.1	1.0	1.8	0.9	1.4
Q5/Q1 non-agr	0.7	14.5	8.1	6.6	0.8	0.8	1.4

lower quintiles, agricultural households show a lower per capita income than non-agricultural ones. Conversely, in the higher one, agricultural households show an average per capita income higher than non-agricultural.

As expected agricultural households earn the largest part of net income from farming (about 60 % of total). The share of the richest among agricultural households is over 44 % of total: a figure that should be read together with their small number (5.3 %). Overall, the 10 % of families included in the higher quintile (agricultural and non-agricultural) earn more than 50 % of total income from farming.

Another good example of the potential utility of the new dataset is the analysis of the distribution of support from sector policy among different household groups. The Single Farm Payment (SFP), a direct transfer decoupled from the level of farm production, is the most important measure within the EU Common Agricultural Policy, in supporting farmers' income. The percentage of SFP accruing to each household group is shown in the fourth column. The figures reveal the existence of a distributive bias: the 11 % of agricultural households included in the highest quintile gather more than 50 % of SFP; furthermore for the richest agricultural households about 15 % of total income is represented by SFP. The support contributes to create about 20 % of farm incomes but with some interesting differences among household groups revealing an imperfect targeting of the measure.

The last column shows the average value of a composite well-being indicator including beside income level also information on housing conditions, education, health status and social exclusion.[6] The index is based on new information from the EU-SILC survey assigned to observations included in the FBS sample through the matching procedure. The availability of well-being indicators may represent a powerful tool in enhancing the targeting of agricultural policy. The index shows, as expected, a value increasing with income level; interestingly, the largest part of support from policy accrues to a small group of households with a well-being index well above the average in the total population.

References

D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching. Theory and Practice. Wiley, Chichester (2006)

EU: Regulation Concerning Community Statistics on Income and Living Conditions (EU-SILC), N. 1177, Brussels (2003)

ISTAT: La distribuzione del reddito in Italia. Argomenti n. 38, ISTAT, Roma (2010)

ISTAT: I risultati economici delle aziende agricole. Anno 2008. Statistiche in breve, ISTAT, Roma (2011)

OECD: Handbook on constructing composite indicators. Methodology and user guide. Paris (2008)

[6]The index is the geometric average of a set of class variables; the aggregation through geometric averaging expresses a partial substitutability among different dimensions of well being (OECD 2008).

Pizzoli, E.: Redditi nelle aziende agricole a conduzione familiare. In: Approfondimento, Rapporto annuale sulla situazione del Paese nel 2004, ISTAT, Roma (2005)

Rassler, S.: Statistical Matching : A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York (2002)

Rocchi, B.: Gathering information on total household income within an "industry oriented" survey on agriculture: methodological issues and future perspectives. In: Pizzoli, E. (ed.) Statistics on Rural Development and Agriculture Household Income. Proceedings of 2nd Meeting of the Wye City Group, Rome, 11–12 June 2010, ISTAT, Roma, pp. 521–528 (2010)

Sacco, G.: SAMWIN: a software for statistical matching. Manual, European Centres and Networks of Excellence (CENEX) – Integration of Surveys and Administrative Data (ISAD), Rome (2008)

UNECE, Eurostat, FAO, OECD, World Bank: Rural households' livelihood and well-being. Statistics on rural development and agricultural household income. United Nations, New York (2007)

Outlier Detection via Compositional Forward Search: Application to the Preliminary Data of the 2010 Italian Agricultural Census

Simona Toti, Filippo Palombi, and Romina Filippini

Abstract

Compositions are multivariate variables, whose components are strictly positive and can be interpreted as parts of a whole. An example of composition is obtained by fractionating an agricultural fund into areas grown with different crops. Parts of a composition fulfill a sum constraint, which establishes an implicit relation among them. This form of dependency goes beyond the standard concept of covariance and invalidates the ordinary techniques of statistical analysis. Outlier detection, for instance, becomes a remarkably entangled problem, due to unsuitability of the Mahalanobis metric to describe the distance among different compositional vectors. Among the established algorithms for outlier finding in a given data set, the Forward Search Algorithm admits an elegant extension to the compositional case. We examine such extension and apply the novel algorithm to the preliminary data of the latest Italian Agricultural Census

1 Building Blocks of Compositional Data Analysis

Compositional data analysis has been the subject of a number of papers, pioneered by Aitchison (1986) over the past 20 years. In this section, we review some basic aspects of it, which are well established in the literature. First of all, the Aitchison distance is derived from a normed vector space, built on top of the geometrical simplex in an arbitrary number of dimensions. This structure is then employed in order to introduce the ILR-transform (Egozcue et al. 2003). Next, we discuss the

S. Toti (✉) • F. Palombi • R. Filippini
Istituto Nazionale di Statistica, Via Cesare Balbo 16, 00184 Rome, Italy
e-mail: simona.toti@istat.it; filippo.palombi@istat.it; romina.filippini@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_16,

conditions under which the ILR transformation acts distributionally as a normalizing map on a compositional data set.

1.1 Aitchison Geometry on the Simplex

As a methodology of statistical investigation, compositional analysis finds application in all cases where the main object of interest is a v-dimensional variate x, whose components are strictly positive continuous real variables, to be regarded as portions of a total amount κ. In other words, x belongs to the v-dimensional simplex

$$\mathscr{S}^{(v)} = \left\{ (x_1, \ldots, x_v) : \quad 0 < x_k < \kappa\,, \quad \sum_{k=1}^{v} x_k = \kappa \right\}\,, \qquad v \geq 2\,. \tag{1}$$

The elements of $\mathscr{S}^{(v)}$ are named compositions. Examples of data which fit in with the above definition are easily found in several contexts, ranging from chemical and geochemical data analysis, to balance sheet analysis, partition of agricultural surfaces, etc. The specific value of κ is not relevant in this framework, since different values of κ mark different simplexes, which are, however, in a one-to-one correspondence with each other (for practical purposes, one can always set $\kappa = 1$).

In such circumstances, it is meaningful to consider the components $\{x_k\}_{k=1,\ldots,v}$ only in terms of their relative importance, i.e. with no reference to their absolute size. If this perspective is adopted, comparing different compositions becomes a matter of comparing one by one all the possible pairwise ratios of their components, since these are the only quantities carrying relative information. Along this lines, Aitchison has introduced an appropriate distance function,

$$d_{\mathrm{A}}(x, y) = \sqrt{\frac{1}{2v} \sum_{i,j=1}^{v} \left[\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right) \right]^2}\,, \qquad x, y \in \mathscr{S}^{(v)}\,, \tag{2}$$

where all these ratios enter equally weighted, smoothed, and symmetrized by logarithms. The Aitchison metric looks pretty different from the more popular Mahalanobis one, as it corresponds indeed to evaluating the distance between observations according to a different criterion. The change of metric invalidates most of the methodologies used in classical statistics and imposes the introduction of more sophisticated analytical tools, in order to settle the ground for a mathematically coherent investigation.

Usually, observations admitting a compositional interpretation are not directly comparable, as their parts sum to different constants κ. This kind of situation occurs frequently in practice: think, for instance, of a chemical data set, whose data units consist of the amounts of certain chemical elements (expressed in a given unit) in each of the analyzed samples; obviously, measured values depend on the total

weight of the sample. In order to conform different observations and make them truly consistent, all parts of each sample have to be rescaled by a positive factor, specific to that sample, chosen so as to bring all observations to fulfill the same compositional constraint. This kind of rescaling is mathematically encoded via a closure transformation

$$\mathscr{C}(x) = \left(\frac{\kappa x_1}{\sum_{k=1}^{v} x_k}, \ldots, \frac{\kappa x_v}{\sum_{k=1}^{v} x_k} \right), \qquad x \in \mathbb{R}_+^v \tag{3}$$

which rescales x, though leaving all the pairwise ratios of its components unchanged. It should be remarked that the Aitchison distance is scale invariant, i.e. it only depends on ratios of components belonging to x and y separately. Nevertheless, the closure transformation has a theoretical rôle in the construction of a theory of compositional analysis, as it turns an empirical multivariate data set $\mathscr{D} = \{x^{(1)}, \ldots, x^{(N)}\}$ made of N observations, into a subset $\mathscr{C}(\mathscr{D}) \equiv \{\mathscr{C}(x^{(1)}), \ldots, \mathscr{C}(x^{(N)})\} \subset \mathscr{S}^{(v)}$. In other words, $\mathscr{C}(x) \in \mathscr{S}^{(v)}$, $\forall x \in \mathscr{D}$.

The Aitchison distance can be derived in a natural way by providing the v-dimensional simplex with a vector space structure on the field of real numbers. The sum of two elements of the simplex is here defined by

$$x \oplus y = \mathscr{C}(x_1 y_1, \ldots, x_v y_v), \qquad x, y \in \mathscr{S}^{(v)}, \tag{4}$$

and the product of an element of the simplex by a scalar is defined by

$$\alpha \otimes x = \mathscr{C}(x_1^\alpha, \ldots, x_v^\alpha), \qquad \alpha \in \mathbb{R}, x \in \mathscr{S}^{(v)}. \tag{5}$$

Note that the closure $\mathscr{C}$ is needed in both Eqs. (4)–(5), in order to project the results of the operations onto $\mathscr{S}^{(v)}$. To complete the vector space construction, an inner product has to be assigned. This is given by

$$\langle x, y \rangle_{\mathrm{A}} = \frac{1}{2v} \sum_{i,j=1}^{v} \ln\left(\frac{x_i}{x_j}\right) \ln\left(\frac{y_i}{y_j}\right). \tag{6}$$

Proofs that Eqs. (4)–(6) fulfill the standard algebraic properties of vector sum, scalar multiplication, and inner product are straightforward and will not be reproduced here. Instead, we conclude this section by recalling that once a vector space is provided with an inner product, a metric structure can be automatically inferred. The norm induced by the inner product is indeed given by $||x||_{\mathrm{A}} = \sqrt{\langle x, x \rangle_{\mathrm{A}}}$, and the distance induced by the norm is accordingly defined by

$$d_{\mathrm{A}}(x, y) = ||x \ominus y||_{\mathrm{A}}, \qquad x, y \in \mathscr{S}^{(v)}, \tag{7}$$

where $x \ominus y \equiv x \oplus [(-1) \otimes y]$. The equivalence of Eqs. (2) and (7) qualifies the Aitchison distance as the result of a metrization procedure operated on the

v-dimensional simplex and gives compositional analysis a high degree of internal coherence.

1.2 Isometric Logratio Transformation

Having closed empirical data via the mapping $\mathscr{D} \to \mathscr{C}(\mathscr{D})$, one is left with a set of compositions, described as tuples belonging to a v-dimensional simplex. Though legitimate, such naïve representation is mathematically redundant, since the sum constraint makes the tuple's components dependent on each other. More generally speaking, different parts of a given composition $x \in S^{(v)}$ cannot be regarded as independent variables owing to the sum constraint, with the consequence that $\dim(\mathscr{S}^{(v)}) = v - 1$. In order to provide an efficient description of $\mathscr{S}^{(v)}$ as a $(v-1)$-dimensional vector space, we need an isomorphic mapping $g : \mathscr{S}^{(v)} \to \mathbb{R}^{v-1}$, whose image space $\mathbb{R}^{v-1}$ may be interpreted as the domain of some set of truly independent generalized coordinates. Among the attempts which have been examined in the literature, we focus on the ILR. What renders this mapping special is that, besides getting rid of the compositional constraint, it establishes a precise isometry between $\mathscr{S}^{(v)}$ and $\mathbb{R}^{v-1}$, as we are going to see.

The ILR, originally introduced in Egozcue et al. (2003), is defined according to the following procedure. First of all, a set of generators $\{w_k\}_{k=1,\dots,v}$ of $\mathscr{S}^{(v)}$ is introduced,

$$w_k = \mathscr{C}(\underbrace{1,\dots,1}_{k-1 \text{ times}}, \mathrm{e}, \underbrace{1,\dots,1}_{v-k \text{ times}}), \qquad k = 1,\dots,v, \tag{8}$$

The Euler constant "e" in Eq. (8) is placed at the kth position of the v-tuple, which has to be then closed, so as to guarantee that the final result $w_k \in \mathscr{S}^{(v)}$. Any point of $\mathscr{S}^{(v)}$ can be represented as a linear combination of the w_k's,

$$x = \bigoplus_{k=1}^{v} \log x_k \otimes w_k, \qquad x \in \mathscr{S}^{(v)}. \tag{9}$$

Despite its completeness, the set $\{w_k\}_{k=1,\dots,v}$ is not yet a basis of $\mathscr{S}^{(v)}$, in that the number of its members exceeds the faithful dimension of $\mathscr{S}^{(v)}$ by one. Nevertheless, a basis can be obtained from Eq. (8) in a straightforward manner: first, one of the generators is dropped, conventionally w_v, then a Gram–Schmidt orthonormalization process (based on the inner product $\langle \cdot, \cdot \rangle_{\mathrm{A}}$) is performed on the remaining $v-1$ generators. The $v-1$ vectors $\{e_k\}_{k=1,\dots,v-1}$ thus obtained constitute a basis of $\mathscr{S}^{(v)}$. At this point, the ILR-transform of a composition $x \in \mathscr{S}^{(v)}$ is defined as the projection of the v-tuple providing the naïve representation of x, along the basis vector according to the inner product $\langle \cdot, \cdot \rangle_{\mathrm{A}}$, i.e.

$$\mathrm{ilr} : \quad x \in \mathscr{S}^{(v)} \quad \mapsto \quad (\langle x, e_1 \rangle_{\mathrm{A}}, \dots, \langle x, e_{v-1} \rangle_{\mathrm{A}}) \in \mathbb{R}^{v-1}. \tag{10}$$

As previously mentioned, the ILR is an isometry: the Aitchison distance in $\mathscr{S}^{(v)}$ is mapped onto the Euclidean distance in $\mathbb{R}^{v-1}$, i.e.

$$d_{\mathrm{A}}(x, y) = d_{\mathrm{E}}\left(\mathrm{ilr}(x), \mathrm{ilr}(y)\right) , \qquad x, y \in \mathscr{S}^{(v)} , \tag{11}$$

with

$$d_{\mathrm{E}}(u, z) = \sqrt{\sum_{i=1}^{v-1} (u_i - z_i)^2} , \qquad u, z \in \mathbb{R}^{v-1} . \tag{12}$$

A proof of Eq. (11) follows directly from the linearity of the operations introduced in Eqs. (4)–(6). Since the metric invariance is absolutely essential for the present work, we go through the few elementary steps leading to its validation. First, we project x and y along the vectors $\{e_k\}_{k=1,\ldots,v}$,

$$x = \bigoplus_{k=1}^{v-1} x_k^* \otimes e_k , \quad x_k^* = \mathrm{ilr}(x)_k , \qquad y = \bigoplus_{k=1}^{v-1} y_k^* \otimes e_k , \quad y_k^* = \mathrm{ilr}(y)_k . \tag{13}$$

Then, we use the orthonormality relations fulfilled by the basis vectors, on each of the terms coming from a bilinear expansion of the square of the Aitchison distance, i.e.

$$\begin{aligned} d_{\mathrm{A}}(x, y)^2 &= ||x \ominus y||_{\mathrm{A}}^2 = \langle x \ominus y, x \ominus y \rangle_{\mathrm{A}} = \langle x, x \rangle_{\mathrm{A}} + \langle y, y \rangle_{\mathrm{A}} - 2\langle x, y \rangle_{\mathrm{A}} \\ &= \sum_{k=1}^{v-1} (x_k^*)^2 + \sum_{k=1}^{v-1} (y_k^*)^2 - 2 \sum_{k=1}^{v-1} x_k^* y_k^* = \sum_{k=1}^{v-1} (x_k^* - y_k^*)^2 \\ &= d_{\mathrm{E}}(x^*, y^*)^2 = d_{\mathrm{E}}\left(\mathrm{ilr}(x), \mathrm{ilr}(y)\right)^2 . \end{aligned} \tag{14}$$

As far as we are concerned with the detection of compositional outliers, the metric invariance tells us that two compositions lie at large (short) Aitchison distance *iff* their ILR-transforms lie at large (short) Euclidean distance.

1.3 Isometric Logratio as a Distributional Mapping

In the sequel, we shall describe an algorithm aimed at detecting compositional outliers. The identification of one or more outliers in a given data set will be based on a statistical test, whose outcome rests in its turn on a distributional assumption on the data. Though in most applications the simplest statistical hypothesis would involve the normal distribution, it can be easily realized that this is not the case: each component of a given compositional vector is a strictly positive number; as such, the tails of its marginal distribution can never extend to the negative semi-axis, as

would be the case for a normal variable. In fact, Aitchison has shown in Aitchison and Shen (1980) that the most natural distributional assumption for compositional data is the log-normal one. From now on, we shall adopt this view. Accordingly, we shall assume that the observations of our data set $\mathscr{D} = \{x^{(k)} \in \mathbb{R}^v_+\}_{k=1,\dots,N}$ tend to distribute according to an v-dimensional log-normal distribution $\Lambda_v(\xi, \Omega)$ with given log-scale vector ξ and scattering matrix Ω, i.e.

$$H_0 : \quad \{x^{(1)} \sim \Lambda_v(\xi, \Omega)\} \cap \{x^{(2)} \sim \Lambda_v(\xi, \Omega)\} \cap \cdots \cap \{x^{(N)} \sim \Lambda_v(\xi, \Omega)\} . \tag{15}$$

The question arises naturally as to how the ILR acts on $\Lambda_v(\xi, \Omega)$. In other words, if $x \sim \Lambda_v(\xi, \Omega)$, how does $\mathrm{ilr}(\mathscr{C}(x))$ distribute? This question has been partially addressed by Aitchison, who has proved in Aitchison and Shen (1980) the following theorem:

Theorem 1. *$x \sim \Lambda_v(\xi, \Omega)$ iff the closure $\mathscr{C}(x) \sim L_{v-1}(A\xi, A\Omega A^{\mathrm{T}})$, where the rectangular matrix A has dimension $(v-1) \times v$ and is given by $A = [\mathbb{I}_{v-1}, -e_{v-1}]$, $\mathbb{I}_v$ being the unit matrix of order v and e_v being an v-dimensional vector with unit components.*

Here we denote $L_{v-1}(\xi, \Omega)$ the Logistic distribution with p.d.f. given by

$$f(u_1, \dots, u_{v-1}) = \frac{1}{(2\pi)^{v/2}|\Omega|^{1/2}} \left[\prod_{j=1}^{v} u_j \right] \mathrm{e}^{\left\{-\frac{1}{2}[\ln(u/u_v)-\xi]^{\mathrm{T}} \Omega^{-1} [\ln(u/u_v)-\xi]\right\}} , \tag{16}$$

with $u_v = 1 - \sum_{j=1}^{v} u_j$. By using this theorem, it is easy to show that $x \sim \Lambda_v(\xi, \Omega)$ entails $\mathrm{ilr}(\mathscr{C}(x)) \sim \mathscr{N}_{v-1}(\mu, \Sigma)$, i.e. $\mathrm{ilr}(\mathscr{C}(x))$ distributes normally with mean μ and covariance matrix Σ, related to ξ and Ω via precise linear transformations. The converse is also true, i.e. $\mathrm{ilr}(\mathscr{C}(x)) \sim \mathscr{N}_{v-1}(\mu, \Sigma)$ entails $x \sim \Lambda_v(\xi, \Omega)$. In summary, the assumption of log-normality of the original data propagates to the closed data and the ILR-transformed ones according to the distributional chain

$$x \sim \Lambda_v \qquad \Leftrightarrow \qquad \mathscr{C}(x) \sim L_{v-1} \qquad \Leftrightarrow \qquad \mathrm{ilr}(\mathscr{C}(x)) \sim \mathscr{N}_{v-1} . \tag{17}$$

2 The Forward Search Algorithm

In this report we consider the problem of detecting potential outliers within a compositional data set $\mathscr{D} = \{x^{(k)} \in \mathbb{R}^v_+\}_{k=1,\dots,N}$. Among the presently known methodologies that allow to detect and eliminate outlying data, we focus on the one based on the *Forward Search Algorithm* (FSA), originally introduced in Atkinson (1994) and thoroughly discussed in Riani et al. (2009) for the case of multivariate analysis.

2.1 Construction of the signal

The original algorithm applies to the case of normally distributed data, i.e. the null hypothesis assumes that all the observations distribute simultaneously according to $x \sim \mathcal{N}_v(\mu_0, \Sigma_0)$, i.e.

$$H_0 : \quad \{x^{(1)} \sim \mathcal{N}_v(\mu_0, \Sigma_0)\} \cap \cdots \cap \{x^{(N)} \sim \mathcal{N}_v(\mu_0, \Sigma_0)\} . \tag{18}$$

The algorithm makes use of the Mahalanobis distance

$$d_{\mathrm{M}}(x, y|\Sigma) = \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-1} (x-y)} , \qquad x, y \in \mathbb{R}^v , \tag{19}$$

and consists of a recursion of sorting steps. Initially, a subset $\mathscr{S}(m_0) \subset \mathscr{D}$ of m_0 observations is chosen (randomly or close to the bulk of the distribution) and is used in order to provide an estimate $\mu(m_0)$ of the true mean μ_0 and an estimate $\Sigma(m_0)$ of the true covariance matrix Σ_0. Then, the recursion starts and its mth step ($m \geq m_0$) is described as follows:

- $\mathscr{D}$ is sorted according to the increasing values of the Mahalanobis distance $d_{\mathrm{M}}(x)$ of the single observation x from $\mu(m)$, given the sample estimate $\Sigma(m)$ of the covariance matrix, $d_{\mathrm{M}}(x) \equiv d_{\mathrm{M}}(x, \mu(m)|\Sigma(m))$;
- the square of the Mahalanobis distance of the $(m+1)$th observation in $\mathscr{D}$ is stored together with the corresponding identification number and defines the mth value of what is named the signal of the FSA;
- $\mathscr{S}(m)$ is replaced by a new subset $\mathscr{S}(m+1)$, made of the first $m+1$ observations of $\mathscr{D}$;
- a new estimate of μ_0 and Σ_0 is obtained from the observations in $\mathscr{S}(m+1)$. Since the observations in $\mathscr{D}$ are ordered, the estimate $\Sigma(m+1)$ of Σ_0 is biased and must be corrected in order to remove the effects of truncating $\mathscr{D}$ at the $(m+1)$th observation. The correction is based on a formula first derived by Tallis in (1963).

The sorting procedure repeats iteratively until the estimate of the distributional parameters involves all the inliers and the occurrence of the first outlier shows up as a break-point (or a jump) in the values taken by the signal as a function of m.

2.2 Testing the signal

In order to give a quantitative meaning to the break-point and to associate it with the occurrence of an outlier, a statistical test, relying on the normality hypothesis, has to be performed. At each step of the recursion, the computation of the signal is accompanied by a measurement of its theoretically admissible lowest and highest values at $(1-\alpha) \times 100\,\%$ confidence level under H_0. The envelope of the pointwise values of these lower and upper thresholds for $m_0 \leq m \leq N-1$ generates two curves that surround the signal until an outlier enters the estimate of the

distributional parameters. The violation of the envelopes signals the income of an outlier within $\mathscr{S}(m)$. The computation of the thresholds is made possible by the fact that the square of the Mahalanobis distance from the mean ideally distributes as a χ^2_v variable under H_0. When the data set is sorted as prescribed by the FSA, the Mahalanobis distances from the mean constitute a set of v order statistics. The distribution of the signal, i.e. the $(m+1)$th order statistic, cannot be computed in closed form. Nevertheless, its percentiles (coinciding with the abovementioned α-thresholds) can be obtained from a general result firstly derived in Günther (1977),

$$y_{m+1,N;\alpha} = (\chi^2_v)^{-1}\left(\frac{m+1}{m+1+(N-m)\,f_{2(N-m),2(m+1);1-\alpha}}\right), \tag{20}$$

where the symbol $f_{a,b;\alpha}$ denotes the α-percentile of the Fisher distribution with parameters (a, b), and $m_0 \leq m \leq N-1$.

3 Extension of the FSA to the Compositional Case

In order to be applied to a compositional data set, the FSA has to be modified in two respects:

- The construction of the signal has to be performed with the Mahalanobis distance being replaced by the Aitchison one. There is a practical advantage in implementing this step. While the measurement of the Mahalanobis distance at the mth step rests on the mth estimate of the covariance matrix, this is not the case with the Aitchison distance. If the covariance matrix is estimated with a finite level of precision, the Mahalanobis distance is affected by a statistical uncertainty which is not present in the Aitchison distance. In other words, the statistical fluctuations in the estimate of the covariance matrix turn virtually the value of the Mahalanobis distance into a spectrum of values.
- The statistical tests, i.e. the construction of the envelopes has to take into account the distributional changes in the null hypothesis. We have seen in Sect. 1 that a convenient choice for the distributional hypothesis of a compositional multivariate is expressed in terms of the log-normal distribution. Accordingly, the statistical tests performed on the signal have to be modified so as to compare the value of the $(m+1)$th ordered Aitchison distance with the thresholds corresponding to the α- and $(1-\alpha)$-percentiles of its distribution.

In order to obtain the percentiles of the distribution of the $(m+1)$th ordered Aitchison distance, we use the isometric property of the ILR. Under the null hypothesis, the observations of $\mathscr{D}$ are log-normally distributed. This entails that the observations of $\mathrm{ilr}(\mathscr{C}(\mathscr{D}))$ are normally distributed. Moreover, the Aitchison distance between two observations of $\mathscr{D}$ equals the Euclidean distance between the corresponding ILR-transformed data, as shown in Eq. (11). It follows that the distribution of the Aitchison distance under the null hypothesis coincides with the distribution of the Euclidean distance under the normality hypothesis

in the Euclidean space obtained from the ILR transformation. The distribution of the squared Euclidean distance under the hypothesis of normally distributed multivariate data has been studied in Ruben (1962). Here it is shown that the cumulative distribution function $H(d^2; \mu, \Sigma)$ of the squared Euclidean distance $d^2 = d^2_{\mathrm{E}}(x, \mu)$ of a normally distributed variable $x \sim \mathcal{N}_v(\mu, \Sigma)$ from the mean μ, can be expressed in terms of an infinite sum of distributions of χ^2-type. More precisely, if $F_v(d^2)$ denotes the cumulative distribution function of χ^2_v, a central chi-square with v degrees of freedom, then $H(d^2; \mu, \Sigma)$ is given by

$$H(d^2; \mu, \Sigma) = \sum_{j=0}^{\infty} c_j(\mu, \Sigma, p) F_{v+2j}(d^2/p), \tag{21}$$

where the coefficients c_j depend on the distributional parameters and can be recursively obtained, and p is a scale factor that can be adjusted to improve the convergence properties of the series. A description of the recursion is beyond the aims of the present report and is detailed in Palombi et al. (2014), as well as some technical aspects of the numerical implementation of Eq. (21). Here, we just observe that the infinite sum can be truncated with no uncontrolled systematic error and that on average a very good approximation is given by the first few terms of the expansion. This leads to a fast evaluation of $H(d^2; \mu, \Sigma)$, and consequently of the envelopes of the FSA. It should also be noticed that $H(d^2; \mu, \Sigma)$ depends on the distributional parameters μ and Σ, i.e. it is not a pivotal distribution. Accordingly, the advantage of having a clean signal, i.e. a signal which does not depend on the estimate of the covariance matrix, is compensated by the disadvantage of having thresholds that depend upon such estimates. Moreover, the sample estimate of Σ obtained from $\mathscr{S}(m)$ suffers from a different bias than in the original formulation of the algorithm, since the truncation of the sample normal distribution obtained out of the ILR-transformed data is not operated here on an elliptical surface, but instead on a spherical one. Correcting this bias requires a different procedure than described in Tallis (1963). The new procedure is detailed in Palombi et al. (2014).

4 A Case Study: The 2010 Italian Agricultural Census

As an example, we apply our compositional FSA to the preliminary data of the sixth Italian agricultural census, held on October the 24th 2010. Agricultural holdings are stratified by type of crop within each region of Italy. Here we focus on the province of Alessandria and the local holdings growing three cereals: soft and spelt wheat, barley, and corn. Our data set is made of $N = 148$ trivariate observations, normalized so that $\kappa = 1$. Inverse fractional parts are shown against each other on a log–log scale in scatter-plots (a)–(c) of Fig. 1. Logarithmic scales are needed in order to emphasize data with extremely low values of one or more parts. Out of the N observations, the FSA detects seven outliers, shown as circled points.

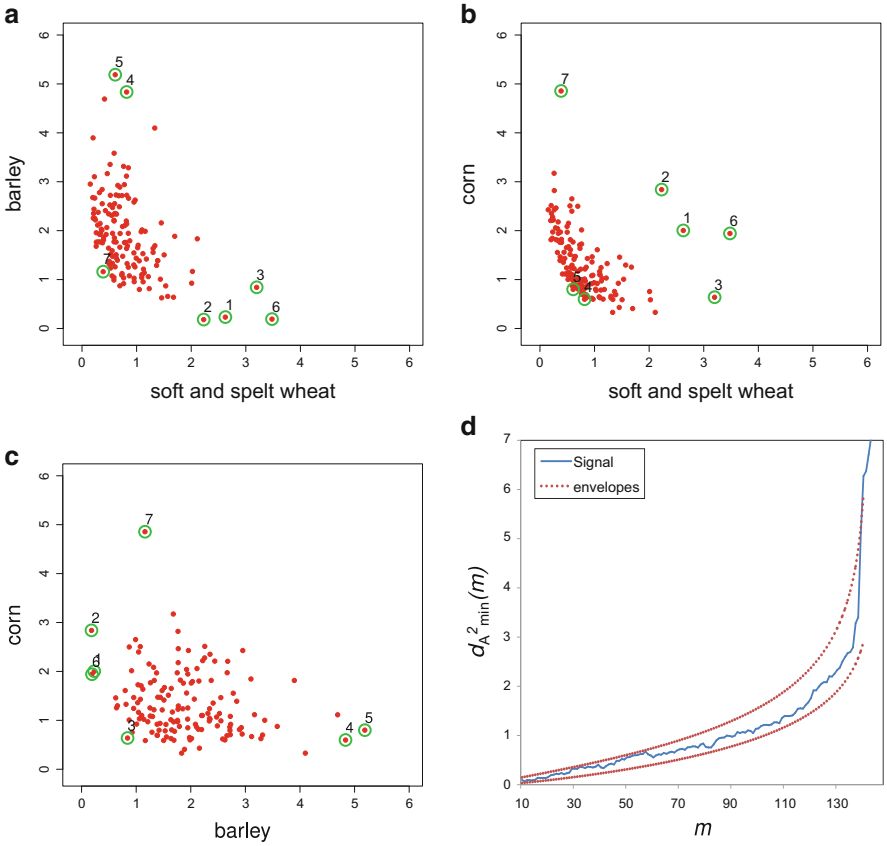

Fig. 1 Results of the compositional FSA on the agricultural data: (a)-(c) Scatter-plots on a log-log scale of the inverse fractional parts against each other; (d) Forward signal and the 1 and 99% over-imposed envelopes at m = 142

Plot (d) of Fig. 1 shows the forward signal and the 1 and 99 % over-imposed envelopes at $m = 142$ (see Riani et al. 2009 for an explanation of the meaning of the over-imposed envelopes). Though not far from the bulk of the distribution, the units marked as outliers are statistically incompatible with the distribution of the remaining data.

References

Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1986)

Aitchison, J., Shen, S.M.: Logistic-normal distributions: some properties and uses. Biometrika **67**(2), 261–272 (1980)

Atkinson, A.C.: Fast very robust methods for the detection of multiple outliers. J. Am. Stat. Assoc. **89**, 1329–1339 (1994)

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelò-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geology **35**(3), 279–300 (2003)

Günther, W.C.: An easy method for obtaining percentage points of order statistics. Technometrics **19**(3), 319–321 (1977)

Palombi, F., Toti, S., Filippini, R.: Numerical Reconstruction of the Covariance Matrix from a Spherically Truncated Multinormal Distribution. ArXiv e-prints, 1202.1838v2 (2014)

Riani, M., Atkinson, A., Cerioli, A.: Finding an unknown umber of multivariate outliers. J. R. Stat. Soc. B **71**(Part 2), 447–466 (2009)

Ruben, H.: Probability content of regions under spherical normal distributions, IV: the distribution of homogeneous and non-homogeneous quadratic functions of normal variables. Ann. Math. Stat. **33**(2), 542–570 (1962)

Tallis, G.M.: Elliptical and radial truncation in normal populations. Ann. Math. Stat. **34**(3), 940–944 (1963)

Innovative Approaches to Census-Taking: Overview of the 2011 Census Round in Europe

Paolo Valente

Abstract

In the course of the year 2011, almost all European countries have conducted the population and housing census. About half of the countries in Europe conducted the 2011 census using an alternative methodology to the traditional census, in most cases for the first time. In general, the alternative methodologies adopted are based on the use of data from registers, either as the only source of census data, or in combination with other data sources. There are also innovative methods that do not make use of registers, like the French "rolling census". This paper discusses the reasons that pushed many countries to consider alternative census methodologies. An overview of the different alternative approaches to census-taking developed in Europe is presented, with an attempt to evaluate the implications in terms of data quality, costs and organization.

1 Introduction

The population and housing census can be considered as one of the "pillars" of national statistical systems. In fact, the census provides the benchmark for the population count at national and local levels, and yields information on the characteristics of the population at fine levels of territorial detail and for small population

The views expressed here are those of the author and do not necessarily reflect the views of the United Nations.

P. Valente (✉)
United Nations Economic Commission for Europe (UNECE), Geneva, Switzerland
e-mail: paolo.valente@unece.org

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_17,

groups on which data cannot be collected using sample surveys. Furthermore, the census allows the creation of a solid frame that can be used to draw samples for all surveys conducted by the National Statistical Institutes and other relevant statistical institutions (UNECE 2006).

The population census plays a key role in all countries, but it is particularly important in countries where vital statistics are not complete and accurate, as the census is the only statistical source providing in these countries accurate and detailed estimates of the population size and structure.

In July 2005 the United Nations Economic and Social Council (ECOSOC) adopted a resolution on the 2010 World Population and Housing Census Programme (UN-ECOSOC 2005), in which Member States were urged to carry out a population and housing census at least once in the period 2005–2014, and to disseminate the census results in a timely manner. Similar resolutions were adopted by ECOSOC in connection with previous census rounds.

At the global level, the situation for the 2010 census round seems to be better compared to the 2000 round. The United Nations Statistics Division (UNSD), which is responsible at the global level for the 2010 World Population and Housing Census Programme, released in March 2011 a report (UNSD 2011) according to which it was expected that by the end of the census round in 2014 at least 227 countries and areas (covering approximately 99 % of the world's population) would have conducted at least one census. Only six countries or areas had not indicated a planned date for their 2010 round census. This is a significant improvement compared to the 2000 census round (covering the period 1995–2004) when 26 countries or areas—mainly in Africa—did not carry out a census.

According to the report by UNSD, in Europe it was expected that virtually all countries would conduct at least one population census by 2014. The large majority of countries in Europe have conducted the census in 2011, which is the reference year for the European census programme (which means that EU member states were expected to conduct their census in 2011). The report also shows that, compared to other continents, Europe stands out for the high number of countries that conducted the census using data from registers or adopting other methods alternative to the traditional census.

This paper presents an overview of the 2010 census round in Europe, focusing on the census methods used in the different countries. In the next sections, the main census methods are presented, starting with the "traditional census" and the problems that countries may face using this approach. The main alternative census methods are presented in the following sections. For each of them, relative advantages and disadvantages are discussed, together with some implications on census organization, costs, data quality and coverage. Finally, information is presented on which census methods have been adopted by the various countries across Europe for the census of the 2010 round, together with a comparison with the methods used for the previous (2000) census round.

2 The Traditional Census Approach

For many centuries, since the censuses taken in ancient Egypt, the methodology used for the census has been basically the same, consisting of the direct count of all individuals and their characteristics through the completion of population lists or—more recently—census forms. This information is collected in the field across the whole country in a relatively short period of time, normally lasting a few weeks.

In the traditional census there are two alternative methods of enumeration. In the first method the census enumerators are responsible for collecting the information from the households during an interview and completing the forms. This approach is usually adopted in countries where a relatively high proportion of the population has minimal education or is illiterate. This method is particularly expensive because enumerators have to conduct interviews in addition to delivering and collecting the forms.

In the second method, the enumerators deliver the forms to the households and collect them some days later when they have been completed. The forms are filled in by the members of the household, normally by a designated member called the reference person (indicated in some countries as "head of the household"). In some countries where this method is adopted, the postal system is used instead of the enumerators for the delivery and/or collection of the forms.

2.1 Problems Associated with the Traditional Census

Although the concept of the traditional census is relatively simple, its implementation is a huge and very complex operation that requires significant financial resources, the participation of various administrations at the central and local levels, and the recruitment and training of a large work force to be employed on a temporary basis as census field staff (enumerators, supervisors, etc.).

From the point of view of the *census management*, there are a number of problems and issues to be faced when the traditional approach is adopted, including the following:

1. *Very high cost*: the census conducted in a traditional way is very expensive. The main cost item is for the temporary work force (enumerators, supervisors, etc.) that has to be recruited and trained, and has to work for a few weeks or longer periods. Considering that an enumerator is needed on average for about 100 households (this is a rough estimate, as the real number depends on the characteristics of the census, the territory and other factors), it is clear that the number of persons to be recruited is very high, and so is the cost. Apart from the cost for census field staff, the cost of printing, distributing, collecting a huge number of census forms, entering the data (manually or using scanners) and processing them is also very high. An analysis of data from the 2000 round of censuses conducted by United Nations Economic Commission

for Europe (UNECE) showed that a traditional census could cost as much as about US$20 per capita in purchasing power parity (ppp) units (United Nations 2008, pp. 39–41).
2. Not only the very high costs, but also the *cost distribution over time* and in particular the peak around the period of the fieldwork can create problems for the management of the traditional census.
3. In many countries, it is difficult to *recruit a large number of temporary census staff* for the fieldwork operations, taking into account that they must have the necessary skills but can be employed only for a short period.
4. In many National Statistical Institutes, once the census operations have been completed, it is not possible to *retain the staff that worked for the census*; they are often reallocated to other services or released. In this case, the knowledge accumulated while planning and conducting the census is lost unless the same staff can be re-employed for the next census.

From the point of view of the *organization of the fieldwork operation*, there are also problems associated with the traditional census, including:

1. The *cooperation of various administrations* at the national and local level is normally necessary to conduct an operation as complex as the traditional census; this may pose problems in some countries, especially if the budget does not fully cover the census expenses, or if the respective tasks and responsibilities of the various administrations involved are not clearly specified.
2. There are increasing *difficulties to enumerate certain population groups*, particularly those characterized by high mobility and multiple residences (including young professionals, students, workers, retired people, or other categories who commute regularly between two or more places). In general, it can be difficult to find these persons at home in order to fill in the census forms. Moreover, identifying the place of usual residence for these people is often complicate. A partial solution to this problem is the possibility for the respondents to complete the census forms on the Internet, which is offered as an option by an increasing number of countries.
3. In many countries, an increasing *reluctance of the population to participate in the census* has been observed over the last years. This can be due to various reasons, including: reluctance to open the door for security reasons, in particular by old people or in areas with security problems; distrust towards the statistical institutes or more in general the public authorities; fear that the information collected could be used for purposes other than the statistical use; reluctance to provide information that is already available in registers or other administrative sources.

Finally, there are also some problems with the *outputs* produced by the traditional census, including:

1. The *timeliness of the census results* is often an issue at least for certain categories of users of the traditional census, because the results are normally available a relatively long time after the data collection, due to the need to process a huge amount of material and information.

2. The *frequency of the results* may also not be sufficient for certain categories of users who need "fresh" data regularly updated: for these users, updates only every ten years are not sufficient.
3. The *information content is limited* by the characteristics of the enumeration, in particular when the forms are completed by the respondents. The number of questions and the time necessary to complete the forms must be limited, and questions that may be complex or potentially sensitive for the respondents have to be avoided.

2.2 A Variation of the Traditional Census: The Use of Long and Short Forms

In order to address some of the shortcomings of the traditional census, a possible solution consists of using two different forms: a long form is used to collect detailed information from a sample of the population, while a short form is used for the majority of the population, to collect only very general information used for the population count. This approach has been used for instance in the United States and Canada since the 1970s.

This method has the advantage of providing extensive information on the characteristics of the population (from the long form), and at the same time reducing substantially the amount of information collected and processed, and limiting the complexity and costs of the census operations. On the other hand, the information present in the long form is available only for a sample of the population, and therefore the information detail is limited both for small areas and for small population groups.

For the 2010 US census, the long form has been replaced by a large household sample survey (the American Community Survey, or ACS) that is conducted every year and provides detailed demographic, social and economic data about households. As a result, the new US census model is based on a decennial traditional enumeration—conducted in 2010 using only a short form—with yearly updates of the population characteristics on a sample basis provided by the ACS.

3 The Register-Based Census

Starting in the 1970s, some Nordic countries began working on a totally different approach to the census, where the traditional enumeration was replaced by the use of administrative data coming from various registers (population register, cadastre, social security, etc.) through a matching process, making use of a personal identification number. This approach, adopted for the first time in Denmark in 1981, permits the production of census data at a limited cost and with relatively limited work, once a good quality system of statistical registers has been set up. This approach has the advantage of placing no burden on individuals, and data

are potentially available every year. Moreover, there is no cyclic distribution in the costs and census staff, as they are distributed relatively evenly across time. It should be noted, however, that setting up and maintaining a statistical system based on registers requires important initial investments and a very long development time (UNECE 2007). Moreover, this approach requires good cooperation between the statistical institute and the authorities responsible for the registers, legislation which allows using register data for statistical purposes and matching records across registers, and finally the acceptance by the public of such a system. All these conditions are met in all the Nordic countries, which adopted this approach in 2011.

A disadvantage of this approach is that the characteristics to be collected are limited to those available in the registers, and the quality of the data produced is dependent on the coverage and quality of the registers themselves. Statistical agencies, however, can combine data from different registers to assess and increase quality and derive new variables. Statistical agencies are also dependent on register authorities (see the requirements listed above), but in the Nordic countries in general there is good cooperation. Establishing and maintaining a high quality register-based statistical system requires significant resources and societal will. However, once such a system is set up, it can be used to efficiently produce a wide range of statistics in addition to census data.

4 The "combined census", Based on Data from Registers and Other Sources

Many countries have population and other registers that potentially could be used for the census, but the coverage and data quality are not sufficient for complete reliance on these registers to produce census data, or some key census variables are not available. Some of these countries in the last years decided that they can still use register data and integrate them with data from other sources in order to produce the census results. Different approaches to this "combined census" exist, depending on what other data sources are used, and how they are used in combination with the register data. Some of these approaches are presented in this section.

4.1 Combining Data from Registers and Existing Surveys

A first approach to the combined census consists in using the results from existing household surveys in combination with register data. An example is the so-called Virtual census conducted in the Netherlands in 2001 and 2011, where register data are integrated with results from the labour force survey (LFS) in order to produce census data. The Netherlands decided to develop this method because they could not obtain from the registers all the necessary information for some of the economic characteristics. Therefore, information on these characteristics is derived based on results from the LFS.

A necessary prerequisite for implementation of this approach, as for the register-based census, is the capacity to link information from different sources at the unit record level. As this method does not require a field data collection, there is no respondent burden on households, and the costs are relatively limited. Moreover, census results are consistent with survey results for common variables. However, the processes to successfully link information on individuals from registers and surveys, and to produce information on households are quite complex.

4.2 Combining Data from Registers and an Ad-hoc Survey

A variation of the previous approach is to combine data from the registers with data from a sample survey conducted ad-hoc for the census. The survey is conducted to evaluate the accuracy of the population or address registers and to collect information on topics that may not be covered in registers, or for which the coverage and quality of registers is not sufficient. The method has the advantage of testing the accuracy of the population register and consequently being able to adjust population counts derived from it. This method was adopted in 2008 by Israel, and in 2011 by other countries including Poland, Spain, Switzerland and Turkey.

4.3 Combining Data from Registers and Full Enumeration

Some countries decided to conduct a census in which the enumeration is based on data from registers, but there is still a full field collection of characteristics on all individuals. This enables variables not available in registers to be obtained in the field as well as providing information about the accuracy of the population count based on registers. This approach is more expensive than the previous ones (presented in Sects. 4.1 and 4.2) because of the full field enumeration. But it is in general less expensive than a traditional census, because of efficiencies in field operations made possible by the use of register data. Compared to a register-based census, this method is clearly much more expensive and poses response burden on the public, but on the other hand it provides improved precision of the results and may help improve the coverage and quality of the registers. For this reason, this approach is often selected for the transition period from a traditional to a register-based census. A significant number of countries in the European Union used this approach for the 2011 census (see Sect. 6).

5 The Rolling Census

Some countries do not have population registers, and therefore cannot adopt the methods presented above. Some of these countries, however, developed alternative approaches to the traditional census without making use of registers. An original and

very innovative approach was developed in France and it is known as the "rolling census". As the name suggests, under this approach the census is conducted as a cumulative continuous (or "rolling") survey over a long period of time rather than on a relatively short time period. In France a 5-year cycle was adopted for the rolling census, and two different strategies are used for small municipalities (population under 10,000) and large municipalities. Small municipalities are divided into five groups, and a full census is conducted each year in one of the groups. In large municipalities, a sample survey covering 8 % of dwellings is conducted each year. At the end of the 5-year cycle, all the population in small municipalities (amounting in France to about half the total population) is enumerated, and about 40 % of the population in the large municipalities. In total, about 70 % of the country's population is enumerated. This is enough to guarantee robust information at the level of municipality and neighbourhoods, according to the French statistical institute INSEE that developed this method.

The census results are based on rolling averages calculated over the 5-year cycle, and are updated yearly. Since the data collection for the French rolling census started in 2004, the first results for the population at the national level were based on data collected in the 5-year period 2004–2008 and were referred to 2006, which was the central year of the period. This method provides for improved frequency of the data, and spreads out across time the financial and human burden associated with the census. On the negative side, the method can be complex to implement. Complications may arise from the movements of persons across municipalities over the various years. These movements could potentially lead to double counting or to missing certain individuals, although specific mechanisms have been put in place to deal with these cases.

6 Overview of Census Methods Used in Europe

In the previous sections, various methodological approaches to the census were presented.[1] In this section information is presented on the methods that were used by European countries for the census of the 2010 round, based on a survey conducted in 2010 jointly by UNSD and the UNECE. The information is presented separately for the countries that are members of the European Union or the European Free Trade Association (EFTA) and for the other European countries that participated in the survey.

[1] A more detailed description of the different census methods is described in more detail in (UNECE 2006), Chap. I and Appendix II. Some examples of implementations of innovative census methods are available on the UNECE website at: http://www.unece.org/stats/documents/2004.11.censussem.htm.

Table 1 Census methods and reference dates for EU and EFTA countries—2010 census round

Country	Census method	Reference date
Austria	Register-based	31 October 2011
Belgium	Register-based	1 January 2011
Bulgaria	Traditional	10 March 2011
Cyprus	Traditional	1 October 2011
Czech Republic	Combined (registers + enumeration)	26 March 2011
Denmark	Register-based	1 January 2011
Estonia	Combined (registers + enumeration)	31 December 2011
Finland	Register-based	31 December 2010
France	Rolling census	1 January 2011
Germany	Combined (registers + enum. + survey)	9 May 2011
Greece	Traditional	16 March 2011
Hungary	Traditional	1 October 2011
Iceland (EFTA)	Combined (registers + survey data)	31 December 2011
Ireland	Traditional	10 April 2011
Italy	Combined (registers + enumeration)	23 October 2011
Latvia	Combined (registers + enumeration)	1 March 2011
Liechtenstein	Combined (registers + enumeration)	31 December 2010
Lithuania	Combined (registers + enumeration)	1 March 2011
Luxembourg	Traditional	1 February 2011
Malta	Traditional	20 November 2011
Netherlands	Combined (registers + survey data)	1 January 2011
Norway (EFTA)	Register-based	19 November 2011
Poland	Combined (registers + survey)	31 March 2011
Portugal	Traditional	21 March 2011
Romania	Traditional	22 October 2011
Slovakia	Traditional	21 May 2011
Slovenia	Register-based	1 January 2011
Spain	Combined (registers + survey)	1 November 2011
Sweden	Register-based	31 December 2011
Switzerland (EFTA)	Combined (registers + survey)	31 December 2010
United Kingdom	Traditional	27 March 2011

Source: Survey conducted by UNSD and UNECE in 2010 and additional information from UNECE

6.1 Census Methods in European Union and EFTA Countries

The data for the 27 EU countries and four EFTA countries (Table 1) show that only 11 countries conducted a traditional census in 2011. The remaining 20 countries (about two thirds of the total) adopted an alternative census methodology.

As the table shows, 12 countries adopted a combined approach, where data from registers were used in combination with a full field enumeration, or with the results of a sample survey (see Sect. 4 above). Among these countries, the most

Table 2 Census methods and reference dates for countries in Eastern and South-Eastern Europe

Country	Census method	Reference date
Albania	Traditional	1 October 2011
Belarus	Traditional	14 October 2009
Bosnia and Herzegovina	Traditional	1 April 2013 (planned)
Croatia	Traditional	31 March 2011
Montenegro	Traditional	31 March 2011
Republic of Moldova	Traditional	2014 (to be confirmed)
Russian Federation	Traditional	14 October 2010
Serbia	Traditional	31 September 2011
The Former Yugoslav Republic of Macedonia	Traditional	31 September 2011
Ukraine	Traditional	2013 (to be confirmed)

Source: Survey conducted by UNSD and UNECE in 2010 and additional information from UNECE

popular approach is using register data in combination with a full enumeration (six countries). As mentioned above, this approach is often adopted by countries that are moving from the traditional census to a register-based census. Among the other countries with a combined census, data from existing surveys have been used together with register data in the Netherlands (where this approach was already adopted in 2001) and Iceland. An ad-hoc sample survey was used together with data from registers in Poland, Spain and Switzerland. Finally, Germany combined data from multiple sources, including registers, a full field enumeration and sample surveys.

Seven countries (the Nordic countries plus Austria, Belgium and Slovenia) conducted a register-based census, while France used the rolling census.

6.2 Census Methods in Eastern Europe

If the majority of countries in the European Union adopted an alternative census methodology for the 2010 census round, the situation is different in Eastern and South-Eastern Europe, where all countries choose the traditional census approach (Table 2). The reason could be that some of the problems associated with the traditional census (see Sect. 2.1 above) do not apply to these countries. For instance, recruiting a large number of temporary census staff could be easier in these countries—compared to countries in Western Europe—thanks to relatively high unemployment, or relatively low labour costs. But there could be also other reasons that make difficult or impossible adopting an alternative census methodology in these countries, like the limited availability of technical or financial resources needed to develop the new census methodology, or the absence of administrative registers of sufficient quality to use the data for census purposes.

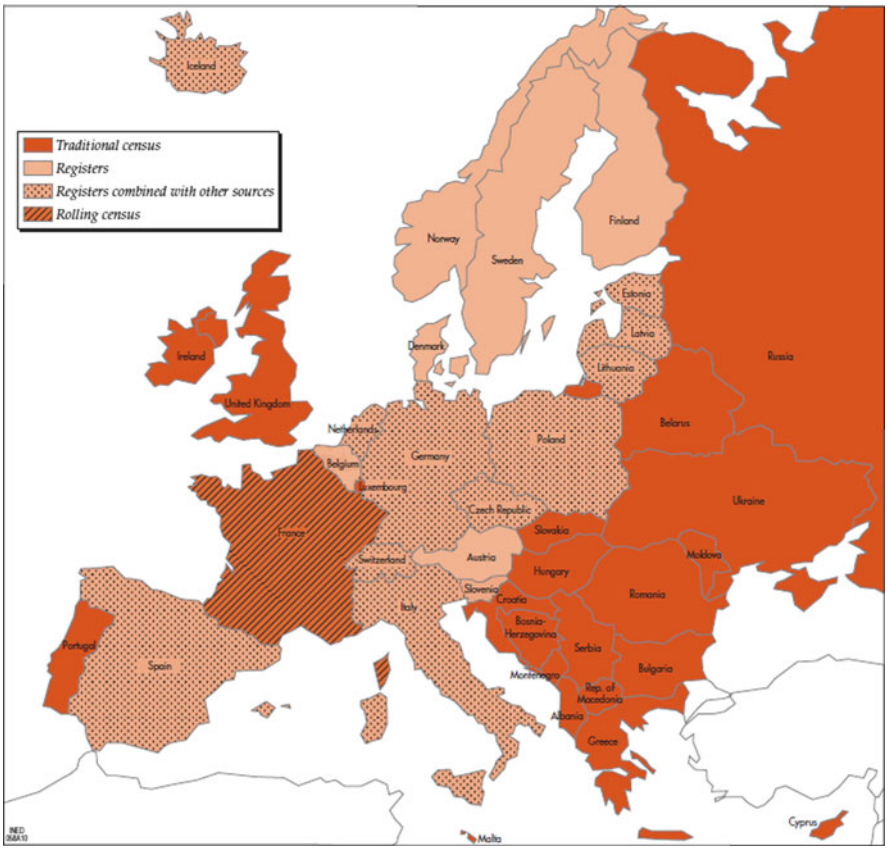

Fig. 1 Methods used by European countries for the 2010 round of censuses. *Source*: Valente (2010), published by INED (modified version)

6.3 The Evolution of Census Methodology in Europe

Based on the information presented above, it emerges that in Europe almost half of the countries (20 out of 41 for which information is available) conducted the census of the 2010 round using an alternative methodology to the traditional census. The majority of them (12 countries) conducted a combined census using data from registers and other sources, seven countries conducted a full register-based census, and France used its original rolling census.

Although a large number of countries in Europe have adopted an alternative methodology, the traditional census is still the most common method in the region, adopted by 21 countries located mainly in Eastern and South-Eastern Europe.

Figure 1 presents the map of Europe by census methodology adopted by countries for the 2010 census round. This figure shows the geographical patterns that were already described above in this article. The traditional census is the

Table 3 Distribution of European countries by census method used in 2000 and 2010 census rounds

Census method					
	2010 round				
2000 round	Traditional	Combined	Register-based	Rolling	Total
Traditional	*20*	5	1	1	27
Combined		*5*	2		7
Register-based			*3*		3
No census	1	2	1		4
Total	21	12	7	1	41

favourite method in all countries in Eastern and South-Eastern Europe, but also in selected EU countries like Ireland, the UK, Portugal and Luxembourg. The register-based approach is adopted in the Nordic countries but also in Austria, Belgium and Slovenia. The combined approach is adopted in many countries in Central Europe, and in Italy and Spain. Finally, France is the only European country where the rolling census is adopted for the 2010 census round.

It is important to note that Europe is the only continent in the world where a significant number of countries conduct the census adopting an alternative methodology to the traditional census. In the rest of the world, virtually all countries use the traditional approach with only a few exceptions, mainly in Asia (registers are used in Bahrain, Israel, Singapore and Turkey).

The trend that sees many European countries moving away from the traditional census and adopting an alternative method started already in the 1970s, as mentioned above, but there was a clear acceleration in the last years. In order to see how many countries changed census method in the last years, Table 3 presents the distribution of European countries by census method adopted in the 2000 and 2010 census rounds.

The values in italics along the main diagonal show that the majority of countries adopted for the 2010 round the same census method as in the 2000 round: the traditional census in 20 countries, the combined census in five and the register-based census in three countries. The fact that five out of seven countries that adopted the combined approach in the 2000 census round have still adopted this approach in the 2010 round and have not moved to a register-based census can be explained by the long development time necessary for a register-based statistical system to be used for the census.

Out of the 27 countries that conducted a traditional census in the 2000 round, seven moved for the 2010 round to an alternative method: it is the combined census for five countries (Czech Republic, Estonia, Italy, Lithuania and Poland), the register-based census for Austria and the rolling census for France. Austria passed directly from a traditional census in the 2000 round to a register-based census in the 2010. This can be considered as a relatively unusual change, but in fact Austria has been working towards the register-based census since long before the census of the 2000 round.

The data presented in Table 3 also show that four countries that had not conducted the census in the 2000 round have actually conducted a census in the 2010 round. Two of them (Germany and Iceland) conducted a combined census, and one (Sweden) a register-based census.

7 Conclusions

With regard to census methodology, Europe can be considered as the world's laboratory, since it is the only continent where a significant number of countries have developed and adopted alternative methods to the traditional census. In the last years, in particular, there was a clear increase in the number of European countries adopting alternative census methods: from 10 in the 2000 census round to 20 in the 2010 round. This trend can be explained by the various shortcomings associated with the traditional census, in terms of costs, management, organization and characteristics of the census outputs. In most cases the alternative census methods make use of data from registers, as unique source of data or in combination with other sources. But innovative approaches have also been developed that do not make use of data from registers, like the French rolling census.

Notwithstanding the trend towards alternative methods, the traditional census was still the most common approach in Europe for the 2010 census round. Probably the traditional census will continue to be the best method for many countries in the years to come, in particular in Eastern and South-Eastern Europe. In fact, although many countries are moving from the traditional census to alternative methods, this is not necessarily the way to follow for all countries. In fact, every method has its strengths and its weaknesses and there is no perfect solution that fits all countries.

Each country needs to decide what will work best in its own context, considering all relevant factors. The most important issue is the quality of the output and its relevance for the uses to which it is put. It is also important that each country clearly documents the method used, evaluates the quality of the census results through established methodologies, and informs the users in a transparent way of possible weaknesses in the data.

References

UNECE: Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing (United Nations), par. 14 on page 5. Available at: http://www.unece.org/stats/publications/CES_2010_Census_Recommendations_English.pdf (2006)

UNECE: Register-Based Statistics in the Nordic Countries – Review of Best Practices with Focus on Population and Social Statistics (United Nations), Ch. 10: Register-based population and housing censuses. Available at: http://www.unece.org/stats/publications/Register_based_statistics_in_Nordic_countries.pdf (2007)

UNECE: Measuring population and housing – Practices of UNECE countries in the 2000 round of censuses. Available at: http://www.unece.org/stats/publications/Publication_on_2000_censuses.pdf (2008)

UN-ECOSOC: Resolution 2005/13. See http://www.un.org/en/ecosoc/docs/docs.shtml (2005)

UNSD: Report on the Results of a Survey on Census Methods used by Countries in the 2010 Census Round. United Nations Statistics Division/Department of Economic and Social Affairs (UNSD/UN-DESA). Available at: http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/docs/ReportOnSurveyFor2010Census.pdf (2011)

Valente: Census taking in Europe – How are populations counted in 2010? In: "Population & Sociétés" no. 467, may 2010 (published by INED). Available at: http://www.ined.fr/fichier/t_publication/1506/publi_pdf2_pesa467.pdf (2010)

Part IV

New Methodological Developments in Economic Studies

Measuring Multidimensional Inequality: Methods and Issues in Empirical Analysis

David Aristei and Bruno Bracalente

Abstract

This paper discusses the main normative measures of multidimensional inequality in well-being and examines the impact of alternative methodological choices in empirical applications. Specific attention is devoted to alternative transformation and normalization criteria, and to normative choices on the degree of substitutability between dimensions and inequality aversion. The empirical application is carried out considering three well-being dimensions (income, health and education) and using Italian data from the IT-SILC survey.

1 Introduction

In recent years, there has been a growing consensus in favour of including other dimensions, beyond monetary indicators, in analysing human well-being and a broad theoretical literature on multidimensional inequality, mainly based on the conceptualization of Sen's (1985) "capability approach", has emerged.

Aggregate measures of well-being and multidimensional inequality require specific assumptions regarding the functional form of the social welfare functions (SWF), the degree of substitutability between well-being dimensions, and the transfer sensitivity (inequality aversion) of well-being between individuals. Alternative assumptions may generate different conclusions on whether inequality increases or decreases over time or is higher or lower across space. Moreover, the results of empirical analysis are sensitive to other methodological choices: the use of a concave (e.g. logarithmic) transformation of income or other dimensions to account for their diminishing returns in the conversion into well-being; the normalization

D. Aristei (✉) • B. Bracalente
Department of Economics, University of Perugia, Perugia, Italy
e-mail: david.aristei@unipg.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_18,

procedures used to aggregate attributes expressed in different units of measurement; the relative weights and the implicit trade-offs between the attributes.

This paper examines some of the main indices proposed in the literature and used in empirical applications. The focus is mainly on measures based on the normative approach to inequality as an extension of the Atkinson–Kolm–Sen univariate approach.[1] Properties and parameter restrictions of such multidimensional measures are outlined and discussed (Sect. 2). The impact of different methodological choices is then examined (Sect. 3). Attention is focused on alternative variables' transformations, normalization criteria and normative choices on the degree of substitutability between dimensions and inequality aversion. To illustrate the impact of measurement issues and methodological choices, an empirical application is performed on Italian data from the 2005 and 2008 IT-SILC surveys, considering three dimensions of well-being commonly used in empirical analyses (Sect. 4). Some remarks conclude the paper (Sect. 5).

2 Measures of Multidimensional Inequality

We assume that the domains of well-being have been identified and that the achievements in all the dimensions are interpersonally comparable. We consider a fixed set of individuals $N = \{1, \ldots, n\}$ (with $n \geq 2$) and a set of $K = \{1, \ldots, k\}$ dimensions of well-being (attributes). A distribution of attributes among the population is an $n \times k$ real-valued matrix X, with the ijth element x_{ij} representing individual i's level of the jth attribute (with $x_{ij} \in \mathbb{R}_+$). Let $\mathcal{D}$ represents the domain of admissible distributions. The ith row of X is denoted $\underline{x}_i = \{x_{i1}, \ldots, x_{ik}\}$ and can be interpreted as a well-being vector for individual i, as it summarizes the achievement of the individual on all the dimensions considered. A multidimensional inequality index $I(X)$ is a continuous real-valued function defined on $\mathcal{D}$ (i.e. $I(X) : \mathcal{D} \rightarrow \mathbb{R}_+$) that summarizes the information on a given distribution matrix. For any $X \in \mathcal{D}$, $I(X)$ measures the degree of inequality and allows ordering distribution matrices in terms of their dispersion.

As in the unidimensional case, each multi-attribute inequality index satisfies, either implicitly or explicitly, a set of properties which define the specific functional form of the index. Following Weymark (2006), the basic properties that a multidimensional inequality index should satisfy can be grouped in two sets: (1) *non-distributional axioms*, which are straightforward generalization of their unidimensional counterparts (*Continuity*, *Anonymity*, *Normalization*, *Replication Invariance*, *Scale Invariance*, *Decomposability* and *Additive Separability*); (2) *distributional properties* (or *majorization criteria*), which provide partial orders that rank distribution matrices in terms of their degree of inequality. In this respect, several authors have tried to provide multivariate generalizations of the

[1] Alternative normative multidimensional inequality indices have been also defined as a generalization of the Gini coefficient. See Decancq and Lugo (2009) for a discussion.

Pigou–Dalton principle of transfer by directly imposing conditions in the space of the distribution matrices X. Kolm (1977) proposed the *Uniform Majorization Principle* (UM), defining the condition that premultiplication of a distribution matrix by a bistochastic matrix (i.e. a non-negative square matrix with row and column sums equal to one) should lead to a socially preferred state. The UM principle imposes that a mean-preserving averaging performed uniformly on all dimensions leads to an increase in social welfare.[2] The UM criterion allows to measure inequality with respect to the dispersion of the multidimensional distribution of the attributes, but fails in addressing the second dimension of multivariate inequality. Atkinson and Bourguignon (1982) argued that a multidimensional inequality index should also account for the dependence between dimensions and developed a dominance criterion, later formalized by Tsui (1999) as the *Correlation Increasing Majorization* (CIM). This criterion is based on the idea that, given two distribution matrices with the same marginal distributions for the dimensions but with different degree of correlation between dimensions, the one with less correlation is socially preferred. Several authors have criticized CIM as it implicitly assumes substitutability between attributes, while as pointed out by Bourguignon and Chakravarty (2003) "there is no a priori reason for a person to regard attributes as substitutes only".

In the literature two approaches have been proposed to define multi-attributes inequality indices. The first, pioneered by Maasoumi (1986) and referred as the two-stage approach, makes the aggregation procedure explicit, by firstly defining a composite well-being indicator for each individual and then applying some univariate inequality index. In the first step, Maasoumi uses information theory to derive a class of aggregation functions with constant elasticity of substitution (CES):

$$S_\beta\left(\underline{z}_i\right) = \left[\sum\nolimits_{j=1}^{k} w_j\left[f_j\left(x_{ij}\right)\right]^\beta\right]^{1/\beta} = \left[\sum\nolimits_{j=1}^{k} w_j z_{ij}^\beta\right]^{1/\beta}. \tag{1}$$

The individual well-being index $S_\beta(\underline{z}_i)$ is a generalized (weighted) mean of order β of the achievements in each well-being dimension. The original values of the indicators in X are generally transformed by applying dimension-specific transformations $f_j(\cdot)$ $(j = 1, \ldots, k)$ to obtain the transformed distribution matrix Z. The dimension weights $w_1, \ldots, w_k$ are assumed to be equal across individuals and sum up to one. The parameter β is related to the degree of substitutability between attributes σ (with $\sigma = 1/(1-\beta)$). Once a composite index of well-being has been defined, overall inequality can be computed by applying a unidimensional inequality measure in the second step. The index proposed by Maasoumi (1986) is obtained by applying a generalized entropy index on $S_\beta(\underline{z}_i)$.[3]

[2] Another multi-attribute Pigou–Dalton transfer principle is the *Uniform Pigou–Dalton Majorization* (UPD) which is less restrictive than the UM principle.

[3] Two-step inequality indices can be also obtained within the normative approach, by specifying an additively separable SWF defined over $S_\beta(\underline{z}_i)$ (see Weymark 2006; Decancq et al. 2009).

Even if the definition of a composite well-being indicator and the application of a univariate inequality index may seem a natural approach, it does not allow to fully capturing the multidimensional nature of well-being. A second approach proposed in the literature is to derive multivariate indices of inequality that satisfy some desirable properties and can be directly applied to the vectors of attributes. Following Weymark (2006), a relative multidimensional inequality measure $I_{\beta}^{M}(Z)$ can be derived starting from a multidimensional SWF $W(Z)$ as the scalar that solves $W[(1-I_{\beta}^{M}(Z))\, Z_{\mu}] = W(Z)$, where Z_{μ} is a distribution matrix in which every observation is replaced by its column mean. Inequality is then the fraction of the overall amount of each attribute that can be given up if every dimension is equalized while keeping the resulting distribution socially indifferent to the original one. Tsui (1995) proposed a multidimensional generalization of the Atkinson–Kolm–Sen approach. After restricting the class of SWF to be continuous, strictly increasing, anonymous, strictly quasi-concave, separable and scale invariant, Tsui obtains an inequality index that can be formulated as (see Brandolini 2008):

$$I^{T}(Z) = 1-\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\prod_{j=1}^{k} z_{ij}^{w_j}}{\prod_{j=1}^{k}\mu_j^{w_j}}\right)^{1-\varepsilon}\right]^{1/(1-\varepsilon)} = 1-\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{S_0\left(\underline{z}_i\right)}{S_0\left(\underline{\mu}\right)}\right)^{1-\varepsilon}\right]^{1/(1-\varepsilon)}, \tag{2}$$

where μ_j is the mean of attribute z_j, $S_0(\underline{z}_i)$ is a special case of the Maasoumi individual composite well-being index $S_{\beta}(\underline{z}_i)$, with $\beta = 0$ (i.e. Cobb–Douglas function), $S_0(\underline{\mu}_i)$ is the corresponding composite well-being of the "average" individual (i.e. the individual endowed with the mean value of each attribute) and the parameter ε (with $\varepsilon \geq 0$) reflects social aversion to inequality.

The Cobb-Douglas aggregation function and the *strong ratio-scale invariance* assumption of the Tsui index were questioned by Bourguignon (1999), who proposed a multidimensional inequality index based on a more flexible CES function, which also depends on the inequality aversion parameter:

$$I^{B} = 1 - \frac{1}{n}\frac{\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\left(w_j z_{ij}^{\beta}\right)^{1/\beta}\right]^{1-\varepsilon}}{\left[\sum_{j=1}^{K}\left(w_j \mu_j^{\beta}\right)^{1/\beta}\right]^{1-\varepsilon}}. \tag{3}$$

A similar index has been proposed by Decancq et al. (2009), based again on a CES functional form, from which they obtain a generalization of the Tsui index:

$$I_{\beta}^{M}(Z) = 1 - \left[\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\left[\sum_{j=1}^{k} w_j z_{ij}^{\beta} \right]^{1/\beta}}{\left[\sum_{j=1}^{k} w_j \mu_j^{\beta} \right]^{1/\beta}} \right]^{1-\varepsilon} \right]^{1/(1-\varepsilon)}$$

$$= 1 - \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{S_{\beta}(\underline{z}_i)}{S_{\beta}(\underline{\mu})} \right)^{1-\varepsilon} \right]^{1/(1-\varepsilon)}, \tag{4}$$

which is linked to the Bourguignon index [3], since $I^{B} = 1 - (1 - I_{\beta}^{M})^{1-\varepsilon}$.

For the class of multidimensional inequality indices considered, the UM principle is satisfied for $\varepsilon > 0$ and $\beta < 1$ while the CIM principle requires that $\varepsilon + \beta > 1$. Then, for the Tsui index CIM is valid only if at least a medium to high aversion to inequality is assumed ($\varepsilon > 1$), while for its generalization [4], as for the Bourguignon index [3], it is satisfied only if the substitutability parameter β is over some level which decreases as the aversion to inequality parameter ε rises.

It is worth noting that, in all the indices considered, the degree of substitution between each pair of attributes is assumed to be the same. This might be an unsatisfactory assumption when working with more than two dimensions. A possible solution is to allow the substitutability parameter to be a function of the achievements, as in Bourguignon and Chakravarty (2003). Decancq and Lugo (2010) also suggest the use of a nested approach by firstly aggregating subsets of dimensions using CES function [1], with a different β for each subset, and then combining these subsets using again the same function.

3 Choices and Issues in Empirical Analysis

The empirical results of multidimensional inequality analysis are sensitive to normative (degree of substitution, inequality aversion, weighting schemes) as well as to methodological choices concerning transformation and normalization procedures.

Transformations of the original well-being dimensions are employed for three main reasons: (1) to capture the diminishing returns in the conversion into well-being of some attributes, especially income; (2) to reduce the effect of extreme values and outliers when the original distribution is markedly skewed; (3) to rescale well-being attributes before they can be sensibly aggregated, as they are generally measured in different units. For positive skewed distribution as income, and to capture its diminishing returns, a concave increasing transformation such as the logarithm is often applied. As pointed out by Anand and Sen (2000), the reason for a concave transformation of income "relates to the fact that the valued object ultimately is not income itself, but the things we are able to do with the help of income" and "there is likely to be some diminishing returns in that conversion". However, Alkire and Foster (2010) claim that similar consideration can apply also

to the other components of human well-being, so that "it would be advisable to use a transformation for each component that was substantially the same". Indeed, especially if performed only on some of the well-being components, the log transformation can produce a severe impact on multidimensional inequality and well-being analysis as it flattens inequality profiles of the attributes, modifies the marginal rates of substitution between attributes and changes the correlation structure between dimensions (see Decancq 2011; Aristei and Bracalente 2011).

Transformations of original variables are often performed also to make effective the cardinality assumption for the well-being dimensions, implicit in the multidimensional indices reviewed in Sect. 2. An example is the transformation into a continue variable of the ordinal scale perception of individual health by means of the fitted values of an ordered probit or logit regression of self-assessed health status on some objective health-related variables and other socio-demographic characteristics (van Doorslaer and Jones 2003).

In conjunction with cardinality, other implicit assumptions of multidimensional well-being and inequality indices are the commensurability of well-being dimensions and their ratio-scale measurability. In practice, variables are generally measured in different units, so that some rescaling procedure is necessary to aggregate them in a sensible well-being indicator. A common normalization procedure is the so-called min–max rescaling, given by the linear transformation $z_{ij} = (x_{ij} - m_j)/(M_j - m_j)$, where M_j and m_j are the maximum and minimum values (*goalposts*) associated with the jth variable. As suggested by Anand and Sen (2000), to make meaningful comparisons across time and space possible, the goalposts should be fixed and predetermined. One limit of this rescaling is the sensitiveness to the choice of goalposts. Moreover, this linear transformation lacks of coherence with multidimensional inequality indices discussed in Sect. 2, which are scale invariant, but translation sensitive. A better alternative consists in applying a rescaling function $f_j\ (x_{ij}) = a_j\ x_{ij}$, where $a_j > 0$ can be the inverse of a measure of central tendency, as the mean or the median, or the maximum value M_j. As shown by Ebert and Welsch (2004), this normalization is suitable for multiplicative well-being functions (and therefore for the Tsui inequality index), which are strong ratio-scale invariant.

The dimension-weights w_j reflect the relative importance of the attributes. Most well-being indicators rely on the assumption of equal weights, which generally reflects insufficient knowledge or lack of consensus on other alternatives. However, many weighting methods, from data-driven to normative approaches, have been proposed (for a survey, see Decancq and Lugo 2010).

In a multidimensional context, normative choices relate to distributional concerns across dimensions and across individuals. As previously introduced, the parameter β determines the shape of the contours for all pairs of attributes. For $\beta \leq 1$ (i.e. non-negative elasticity of substitution) the well-being indicator is a weakly concave function, reflecting a preference for well-being bundles that are more equally distributed. When $\beta = 0$ the composite well-being function is a Cobb–Douglas, while when $\beta = 1$ it is a linear function of the k attributes and $\sigma \to \infty$ (i.e. attributes are perfect substitutes). As β decreases, society is less willing to compensate

inequalities among dimensions. In the limit, as $\beta \to -\infty$ ($\sigma \to 0$), dimensions are treated as perfect complements and the well-being function is of Leontief type. On the other hand, ε defines societal aversion to inequality and it increases as more weight is given to the lowest part of the distribution of attributes across individuals.[4]

4 Well-Being Inequality in Italy: Some Lessons from Microdata

In this section we assess the impact of alternative normative and methodological choices in an empirical analysis performed on Italian microdata.

4.1 Alternative Choices

To compute multidimensional inequality indices, five assumptions have to be made, relating the weighting scheme, the logarithmic transformation of the monetary dimension, the normalization of attributes, the degree of substitutability and the inequality aversion. For simplicity, we assume equal weights for all the indices considered. We consider two alternative normalizations: (1) a dimension-specific rescaling, dividing the values of each dimension by its respective mean in 2007; (2) the min–max linear transformation, with goalposts fixed to the highest value and to the half of the minimum of each dimension in the 2 years. The choice of the lowest goalpost is due to the necessity of avoiding zero values in the rescaled attributes, which therefore does not exactly vary between 0 and 1. While rescaling by the mean does not modify the distribution of the attributes, the min–max rescaling, implying both scaling and translation, modify inequality assessment when translation sensitive indexes are used. Finally, we compare alternative values of the degree of substitutability and of inequality aversion. In particular, we consider parameter combinations for which multidimensional indices do not necessarily satisfy CIM and allow treating attributes also as complements. We allow β to vary between 1 (perfect substitutability) and -5 (relatively high complementarity), focusing on $\beta = 0$ for which we obtain the Tsui index [2], and ε to vary between 0.3 and 3, a range commonly adopted in empirical analyses.

4.2 Data and Indicators

Data are taken from the 2005 and 2008 cross-sectional waves of the "Statistics on Income and Living Conditions" (IT-SILC) survey and refer to income levels and living conditions at the end of 2004 and 2007 on a sample of individuals aged 16

[4]See Aristei and Perugini (2010) on the effect of assuming homogeneous versus heterogeneous (i.e. country-specific) inequality aversion in comparative analyses of multidimensional inequality.

and more. We focus on the distribution among individuals of three dimensions of human well-being, proxied by equivalized disposable income, an indicator of health status and years of schooling attained.

Equivalized disposable income is obtained by dividing total disposable household income, adjusted with the within-household non-response inflation factor, by the modified OECD equivalence scale. For the aims of our analysis, all members of the same household are assigned the same equivalized disposable income. As inequality indicators are sensitive to the presence of extreme incomes in the tails of the distribution, we decide not to consider negative and zero incomes and adopt a winsorizing procedure for the lowest 0.25% and the highest 0.1% income observations.

Two indicators of health status are alternatively considered. Firstly, we use a categorical variable measuring self-assessed health status. This indicator has the advantage of providing a global assessment of health status, but its subjectivity may limit interpersonal comparability. Moreover, it may lead to underestimate the degree of health inequality (van Doorslaer and Jones 2003). For these reasons, we define a composite cardinal indicator of health status based on the predicted values of an ordinal logit regression of self-assessed health status on three other health-related variables and socio-demographic characteristics.[5] This procedure imposes cardinality, while ensuring that individuals with the same observed characteristics obtain the same health measure (Decancq and Lugo 2009).

The third dimension relates to years of schooling and is constructed by combining information on the highest education level attained and the number of years in post-secondary education. For those individuals who are still in education, we assign a value equal to the years of schooling corresponding to the course currently attended. To avoid computational problems caused by the presence of zeroes, the whole distribution is then translated by one unit.

4.3 Univariate Analysis

Table 1 provides some basic descriptive statistics and shows the evolution of inequality between 2004 and 2007 for the attributes considered. Individual sample weights are used in all the computations presented. Analysing the distribution of the variables rescaled by the mean values, it is possible to note that each dimension in 2007 is generally characterized by less inequality. The logarithmic transformation of equivalized income has, as expected, a great impact in reducing income concentration, but leave the comparison between the 2 years unchanged.

[5]The variables considered include indicators of chronic illness, physical limitations and unmet health treatments, together with income level, years of education, age, gender, marital status and region of residence. All variables are highly significant and have the expected signs. Predicted values are linearly transformed to range from slightly more than 0 (unhealthiest individual) to 1 (healthiest individual).

Table 1 Descriptive statistics and univariate inequality measures (years 2004 and 2007)

	Equivalized income				Health status					
	Income		Log (income)		Self-assessed		Fitted values		Education	
	2004	2007	2004	2007	2004	2007	2004	2007	2004	2007
(a) Descriptive statistics (original variables)										
Mean	16,952	18,102	9.641	9.641	3.631	3.631	0.678	0.670	11.19	11.52
Std. Dev.	11,757	11,170	0.596	0.596	0.895	0.895	0.177	0.187	5.16	5.19
Min	686.7	666.7	6.502	6.502	1	1	0.026	0.020	1	1
Max	141,883	122,900	11.719	11.719	5	5	1	1	23	23
(b) Univariate inequality measures										
(*b1*) *Rescaling by mean values*										
Gini	0.317	0.301	0.035	0.033	0.130	0.128	0.140	0.151	0.258	0.252
ATK1	0.166	0.150	0.002	0.002	0.037	0.039	0.050	0.058	0.145	0.141
ATK2	0.346	0.324	0.005	0.004	0.085	0.093	0.133	0.151	0.387	0.380
(*b2*) *Min–max linear transformation*										
Gini	0.323	0.307	0.053	0.050	0.151	0.148	0.142	0.153	0.270	0.264
ATK1	0.175	0.159	0.005	0.005	0.054	0.058	0.052	0.060	0.177	0.171
ATK2	0.398	0.373	0.011	0.010	0.140	0.157	0.139	0.162	0.535	0.526

Only for the health status variables, we have some changes in the ranking of the two distributions depending on the indicators considered. In particular, when we consider the predicted values of the ordered logit, health inequality appears to be lower in 2004 than in 2007, while when the original self-assessed measure is taken into account inequality remains unchanged. The min–max linear transformation significantly increases inequality in each dimension. This is particularly relevant for the Atkinson inequality indexes and the differences between the two transformations considered tend to increase as the inequality aversion parameter rises. This result is a consequence of the fact that the Atkinson index is scale invariant, but translation sensitive, suggesting that the choice of the transformation functions should be carefully considered.

Table 2 shows the correlation structure between the dimensions of well-being. As it can be noted, correlation coefficients between dimensions are all significant at the 1 % level (and remain stable over the period considered). The logarithmic transformation of income has a minor effect on the correlation structure, leading only to a small decrease in correlation with respect to the fitted values of health status and education in 2008. On the other hand, using the composite indicator of health status instead of the original self-assessed ordinal variable increases correlations between dimensions in both the years. The increase is particularly significant for the correlation between health and education (which passes from 0.4235 to 0.5601 in 2008), while it is moderate for the correlation with respect to income (either in levels or in logs). This suggests that using the composite health indicator will lead to an increase in inequality measured by means of multidimensional indexes satisfying the CIM principle.

Table 2 Correlation structure: pair-wise correlation coefficients (year 2007)

	Equivalized income		Health status		
Variable	Income	Log (income)	Self-assessed	Fitted values	Education
Income	1				
Log (income)	0.8649*	1			
Health status: self-assessed	0.1338*	0.1338*	1		
Health status: fitted values	0.1515*	0.1498*	0.7002*	1	
Education	0.2970*	0.2860*	0.4235*	0.5601*	1

Note: Correlations between self-assessed health status and all the other variables are measured by the Spearman rank correlation coefficient
*Significance at the 1 % level

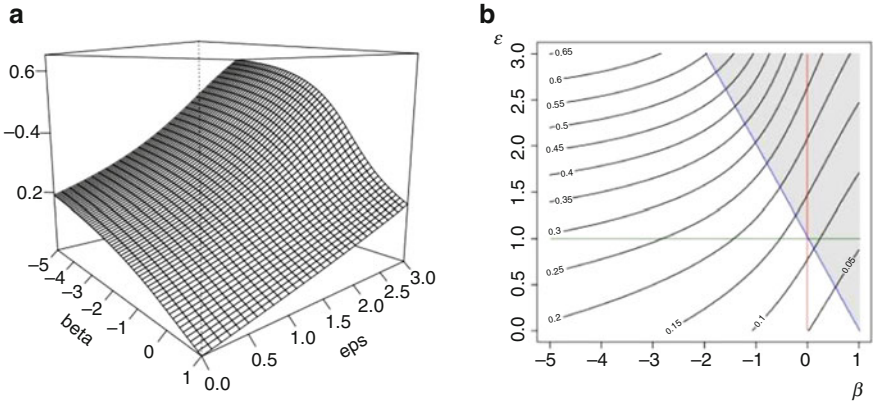

Fig. 1 Multidimensional inequality as a function of β and ε parameters (year 2007). *Note*: the parameter combinations for which IT satisfies the CIM axiom are highlighted in *grey*. (**a**) Surface plot. (**b**) Contour plot

4.4 Multidimensional Inequality Analysis

In Fig. 1 we present the values of the inequality index [4] for 2007 considering alternative combinations of β and ε parameters, using the fitted values of health status and normalizing dimensions by their respective mean values in 2007. Inequality increases as the parameter of inequality aversion rises and dimensions are treated as complements. The contour plot in panel *b* shows that I^M becomes particularly sensitive to changes in the correlation structure between attributes when mid to high inequality aversion is assumed ($\varepsilon > 2$): small decreases in β (i.e. attributes are assumed as more and more complements and society is less willing to compensate inequalities between dimensions) cause increases in multidimensional inequality, moving I^M to higher iso-inequality curves.

Normative choices also affect the pattern of inequality over time. From Table 3 an overall decreasing tendency of inequality can be picked out. Decreases in inequality are however statistically significant only in the area of the parameter space where

Table 3 Changes in multidimensional inequality between 2004 and 2007

	ε			
β	0.3	1	2	3
−2	−0.0106‡	−0.0119‡	−0.0129‡	−0.0130
	(0.0028)	(0.0034)	(0.0053)	(0.0079)
−1	−0.0084‡	−0.0097‡	−0.0094†	*−0.0098*
	(0.0027)	(0.0031)	(0.0045)	(*0.007*)
0	−0.0045	−0.0042	*−0.0031*	*0.0003*
	(0.0025)	(0.0027)	(*0.0033*)	(*0.0045*)
0.7	−0.0019	*−0.0033*	*−0.0011*	*0.0028*
	(0.0027)	(*0.0027*)	(*0.0030*)	(*0.0040*)

Note: Bootstrapped (500 replications) standard errors in parentheses. Daggers ‡ and † denote significance at the 1 % and 5 % levels, respectively. Combinations for which the CIM axiom is satisfied ($\beta + \varepsilon > 1$) are in italics

CIM is not satisfied and the effects of decreasing univariate inequalities and of the increasing correlation between dimensions over time add up, further reducing multidimensional inequality. Negative differences are not statistically significant in the other areas, where the joint effect of increasing correlations and of high inequality aversion compensates the decrease of inequality in each dimension.

Turning to the sensitivity of multidimensional measures to alternative methodological choices, we focus on the Tsui index ($\beta = 0$) with different inequality aversion parameters (Fig. 2). As it can be noted from panel (a), measuring health dimension by means of the estimated composite indicator raises multidimensional inequality, as expected from the univariate analysis. This cardinal measure is characterized by higher inequality than the self-assessed ordinal health status and it leads to an increase in the correlation with the other dimensions that turns into a rise in multidimensional inequality, which is more evident for higher values of ε (as the CIM axiom is satisfied for $\varepsilon > 1$). Concerning the logarithmic transformation of income, it causes a significant downward shift of multidimensional inequality. This gap, which tends to slightly increase with ε, is a consequence of the dramatic decrease in inequality of the monetary component of well-being due to the concave transformation, which becomes much lower than inequality of education and health status. This evidence suggests that, in evaluating multidimensional inequality, the hypothesis of Anand and Sen (2000) is acceptable only if it is sensible to assume that an extremely low level of inequality characterizes the human capabilities proxied by income. It is worth remarking, despite results are not presented, that also normalization procedures are found to significantly affect multidimensional inequality assessment: the min–max normalization leads to higher inequality in all the cases considered, with diverging inequality profiles as ε increases, due to the higher sensitiveness of the min–max transformation to the presence of zeroes.

Turning to changes of multidimensional inequality over time (Fig. 2, panel b), the different assumptions significantly change the observed patterns. When we consider income in levels, using either ordinal health status or the composite indicator, inequality shows a decreasing trend for almost all the values of the inequality

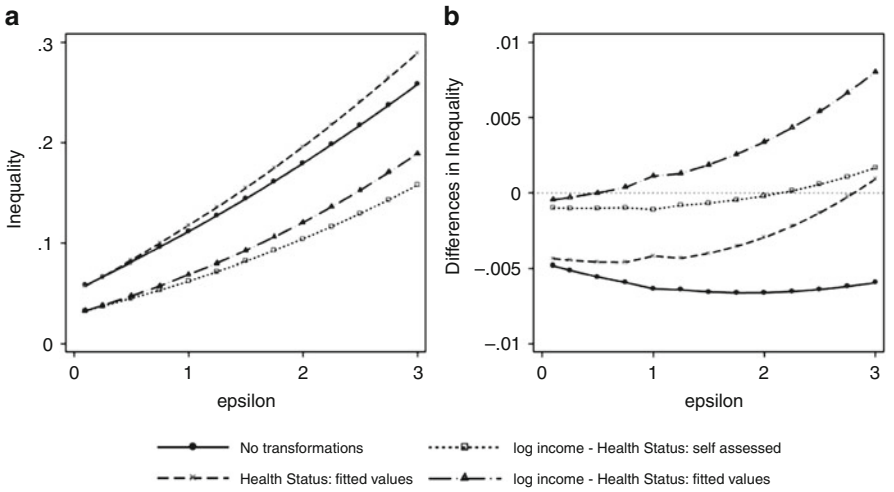

Fig. 2 Tsui multidimensional inequality index. (**a**) Inequality in 2007. (**b**) Differences 2004–2007

aversion parameter considered. The picture is completely different when we consider the logarithm of income. The concave transformation together with the use of the estimated health indicator reverse the ranking of the two distributions: measured inequality in well-being is now higher in 2007 than in 2004 for almost every ε values and the differences are statistically significant for parameter combinations satisfying the CIM axiom. This result is in line with the empirical findings of Decancq (2011) and shows how a controversial (but commonly adopted) assumption as the logarithmic transformation of the monetary dimension may significantly bias empirical results, distorting the picture of the evolution of well-being inequality over time.

5 Concluding Remarks

The empirical findings of our study confirm that normative and methodological choices affect measured multidimensional inequality and show that disregarding these underlying assumptions can be very misleading for the assessment of inequality. Multidimensional indices not only vary with alternative normative choices on the degree of substitution and inequality aversion, but they also depend on attribute transformations and normalization procedures. In this respect, the definition of the indicator of command over resources plays a significant role: the logarithmic transformation of the monetary dimension leads to a flattening of inequality profiles and strongly modifies the implicit trade-offs between attributes. Moreover, the normalization of the attributes is not independent to the properties of the inequality index considered: the commonly adopted min–max linear transformation, besides being sensitive to the choice of the goalposts, is not consistent with the strong

ratio-scale invariance property, while normalization by the inverse of a measure of central tendency is preferable especially for the Tsui index.

It is worth remarking that these aspects are not merely technical issues, but reflect alternative views of social well-being. In particular, the investigation of a multiplicity of normative assumptions on the degree of substitutability between dimensions and inequality aversion, exploring the entire parameters space without imposing the restrictions of the CIM principle, allows to explicit the sources of variation in the results and enriches the assessment of multidimensional inequality in well-being.

References

Alkire, S., Foster, J.: Designing the inequality-adjusted human development index (IHDI). Human Development Research Paper, 2010/28, UNDP (2010)

Anand, S., Sen, A.K.: The income component of the human development index. J. Hum. Dev. **1**(1), 83–106 (2000)

Aristei, D., Bracalente, B.: Measuring multidimensional inequality and well-being: methods and an empirical application to Italian regions. Statistica **71**(2) (2011)

Aristei, D., Perugini, C.: Preferences for redistribution and inequality in well-being across European countries: a multidimensional approach. J. Policy Model. **32**, 176–195 (2010)

Atkinson, A.B., Bourguignon, F.: The comparison of multi-dimensioned distributions of economic status. Rev. Econ. Stud. **49**, 183–201 (1982)

Bourguignon, F.: Comment on 'Multidimensioned approaches to welfare analysis' by Maasoumi. In: Silber, J. (ed.) Handbook of Income Inequality Measurement. Kluver, Boston (1999)

Bourguignon, F., Chakravarty, S.R.: The measurement of multidimensional poverty. J. Econ. Inequality **1**, 25–49 (2003)

Brandolini, A.: On applying Synthetic Indices of Multidimensional Well-Being: Health and Income Inequalities in Selected EU Countries. Temi di discussione, vol. 668. Banca d'Italia, Rome (2008)

Decancq, K.: Global inequality: a multidimensional perspective. Center for Economic Studies – Discussion Paper 11.09. Katholieke Universiteit, Leuven (2011)

Decancq, K., Decoster, A., Shokkaert, E.: The evolution of world inequality in well-being. World Dev. **30**, 11–25 (2009a)

Decancq, K., Lugo, M.A.: Measuring inequality of well-being with a correlation-sensitive multidimensional Gini Index. Economic Series Working Papers 459, University of Oxford (2009)

Decancq, K., Lugo, M.A.: Weights in multidimensional indices of well-being: an overview. Center for Economic Studies – Discussion Paper 10.06, Katholieke Universiteit Leuven (2010)

Ebert, U., Welsch, H.: Meaningful environmental indices: a social choice approach. J. Environ. Econ. Manag. **47**(2), 270–283 (2004)

Kolm, S.C.: Multidimensional egalitarianisms. Q. J. Econ. **91**, 1–13 (1977)

Maasoumi, E.: The measurement and decomposition of multi-dimensional inequality. Econometrica **34**, 991–997 (1986)

Sen, A.K.: Commodities and Capabilities. North Holland, Amsterdam (1985)

Tsui, K.Y.: Multidimensional generalizations of the relative and absolute inequality indices: the Atkinson–Kolm–Sen approach. J. Econ. Theory **67**, 251–265 (1995)

Tsui, K.Y.: Multidimensional inequality and multidimensional generalized entropy measure: an axiomatic derivation. Soc. Choice Welf. **16**, 145–157 (1999)

van Doorslaer, E., Jones, A.M.: Inequalities in self-reported health: validation of a new approach to measurement. J. Health Econ. **22** (2003)

Weymark, J.A.: The normative approach to the measurement of multidimensional inequality. In: Farina, F., Savaglio, E. (eds.) Inequality and Economic Integration. Routledge, London (2006)

Turning the Compulsory Communication Data into a Statistical System

C. Baldi, G. De Blasio, G. Di Bella, A. Lucarelli, and R. Rizzi

Abstract

The Compulsory Communications system is a stream of declarations due by employers to notify the events of activation, termination, extension or transformation of each employment relationship. Thanks to its wide coverage and rich variable set, these data will provide an informative basis: (a) to timely monitor the short-term evolution of the labour market at a very detailed level and (b) to perform complex analysis on the structure of labour demand and labour policies effectiveness. To fully exploit the potential of this administrative source a set of data editing and processing procedures have to be set up, in order to construct and update information on complex statistical units such as employment relationships (jobs). This paper illustrates the building blocks of the current procedure flow and their statistical impact.

1 Introduction

The Compulsory Communications system (*Comunicazioni Obbligatorie*, from now on CO), managed by the Ministry of labour and social policy,[1] is a stream of declarations due by employers to notify the events of activation, termination, extension or transformation of each employment relationship.

[1] See Ministero del Lavoro e delle Politiche Sociali (2010).

C. Baldi • G. Di Bella • A. Lucarelli • R. Rizzi
National Institute of Statistics (Istat), Rome, Italy

G. De Blasio (✉)
Italia Lavoro S.p.A., Rome, Italy
e-mail: GDeBlasio@italialavoro.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_19,

The main objective of the CO, as stated by the legislation, is to provide an information system supporting actions to contrast irregular work but also aimed to implement a database for monitoring and evaluating labour policies. These information must be shared among the main institutions involved in achieving the goals: local authorities, Social security system, Ministry of Labour and social policy (see article 17 Legislative Decree of 10 September 2003 No. 276). It is interesting to note that while there is a strictly administrative purpose, it is also accompanied by an explicit informative/statistical purpose. Two other important legislative steps are the Financial law of 2007 and the Legislative Decree of 30 October 2007 which, respectively, extend the obligation to all the employers including the public administration and make the online transmission compulsory defining official modalities and deadlines, starting from March 1, 2008.

The information contained in CO have a tremendous potential for the analysis of labour market. However to fully exploit this potential many methodological aspects have to be tackled to transform the administrative source into a statistical one.

The scope of this chapter is to illustrate the informative content and the uses of system along with describing the building blocks of the transformation procedures.

The structure of the work is the following. Section 2 describes the scope of the system and statistical uses. In Sect. 3 the main target parameters are presented along with place of these indicators in the labour statistics.

The flow of procedures necessary to pass from the administrative database to the statistical one is illustrated in Sect. 4. Two of the most problematic aspects, the construction of the statistical unit job and the monitoring of data consolidation are presented respectively in Sects. 5 and 6. Some concluding remarks follow.

2 Relevance, Scope and Objectives of the Statistical Compulsory Communication System

The coverage of the system in terms of employers and workers is very wide. More specifically the system covers all public and private employers of all economic sectors (including agriculture, public administration, households as employers) with the exception of armed forces. In terms of workers it includes only regular workers: all employees (including the temporary workers and domestic personnel); part of self-employed workers (project workers, occasional workers, self-employed in entertainment industry, business agents) and stagers.

It excludes regular workers employed in the forces and some managerial contracts of public and private corporations (presidents, CEOs) and obviously irregular workers.

Compared with other administrative sources (e.g. Social Security data), a very distinctive feature of the system is that it allows to measure only flows in and out of employment and not stocks of jobs or persons (see paragraph 3). The communications contain a rich set of variables of the employer (economic activity, workplaces, etc.), of the worker (gender, age, education, citizenship, etc.) and of the job (occupation, type of contract, type of working time, etc.). A statistical system based on this administrative source might greatly enhance the informative support

to policy making and analysis, through two broad classes of uses. (a) Short-term evolution of the labour market monitoring. Since the communication flows into the local and national databases in real time, these indicators can be released very timely. In addition to indicators on the number of activations and terminations (and their difference), statistics on the transformation into permanent jobs, the duration of labour contracts, the cause of termination (voluntary and involuntary) are able to provide a rich picture on the labour market dynamic. Since the system measures all the events happened in the economy, this representation can be very detailed in all the relevant dimensions (geographical, economic activities, characteristics of workers and job). (b) Studies on labour market policies and on the composition of the labour demand. The availability of (longitudinal) microdata opens the possibility to answer a number of questions relevant to policy makers and analysts. Examples are: what is the probability of an agency worker (or other type of temporary worker) to get a permanent job in a certain lapse of time? Are the subsidies devoted to the employability of person with disabilities effective? Which are the occupations most requested by enterprises?

3 Target Parameters and the Relation with Other Sources

The main statistics to track the evolution of the labour market are the number of events related to the employment relationships: activations, terminations, extensions and transformations. These *gross flows* will enrich the understanding of the labour market dynamics by providing the magnitude of the real turnover of jobs which is hidden by the more familiar measures of the net changes. On the other side the system cannot provide reliable measures of stocks, at least for few years to come. A stock of jobs would be in fact limited to those that have experienced at least one event since the start of the system (March 2008). All the employment relationships started before 1 March 2008, with the exception of those which underwent any event of transformation or extension, are unknown to the system. Since these include all those permanent jobs started before that date, the employment stock measured by the system is severely underestimated. However, since sooner or later all the permanent jobs will end, the system will eventually be able to measure stocks as well.

While the system is unable to measure the *level* of the stocks it should in principle provide quite accurate estimates of the *change* of them. This feature derives from the law of motion of the employment stocks:

$$J_t = J_{t-1} + A_{(t-1,t)} - T_{(t-1,t)}$$

That is, the number of jobs at time t (J_t) is equal to the number of jobs at $t-1$ plus the activations $A_{(t-1,t)}$ minus the terminations $T_{(t-1,t)}$ occurred during the period. The net change of jobs is thus simply equal to the difference of activations and terminations from time $t-1$ to t:

$$\Delta J_t = A_{(t-1,t)} - T_{(t-1,t)}$$

Notice that J_t and J_{t-1}, the true number of jobs active in the economy at time t and $t-1$, are not observed by the system, but this is unnecessary to measure their change as long as the system can measure reliably the number of activations and terminations.

The possibility of deriving measures of net changes gives rise to the possibility of comparing these estimates with those traditionally released from other data sources. The two set of indicators deputed to measure the change in employment are the Labour force survey (LFS) estimates and the National accountancy (NA) measures. The main differences between the employment statistics derived from the LFS and the CO system are: (a) the first measures the change in the number of employed person while the second of jobs; (b) the LFS relates to employees in whatever status, while the second measures the jobs of employees and only part of self employment; (c) the LFS statistics are referred to the persons resident in a given area while the most common way to present the CO statistics is in relation to the workplace; (d) the CO system cannot include by definition the irregular work that is present in the LFS. The difference between persons and jobs can be overcome by using the national accountancy data that provide estimates of the jobs. However the differences related to self employment and the presence of irregular work remain.

Another difficulty when CO statistics are compared with the traditional sources arises because of the measures of net changes calculated. While the CO net change in a period is usually measured from the start to the end of the period (e.g. from 1 January to 31 December) the change measured by the LFS or the NA is a change between average stocks (e.g. mean value of year t minus mean value of year $t-1$, or mean value of quarter q minus mean value of quarter $q-4$). A way to make CO statistics more comparable to these other sources is to reduce the extra-noise contained in the start-to-end measure by taking averages of several net changes.

Anastasia et al. (2010) have shown that once the all possible measures are taken to make the statistics as much comparable as possible (e.g. restricting the comparison only to employees, averaging out the net changes over 3 months) the employment dynamics measured with CO and LFS in the Veneto region are quite close, hinting at the fact that the CO estimates are reliable.

4 Data Flow Description

As it is common with administrative sources the phases of data acquisition are not designed and implemented to produce statistics and have to be taken as they are by the statistician.[2] Considering the possible impact on statistical results, however they have to be carefully monitored. For this reason the description of the procedure flow must start from the very first stage of data acquisition and must go through the ETL operations of data storing and filtering.

[2]See Italia Lavoro (2010), Ministero del Lavoro e delle Politiche Sociali (2010).

The system involves a number of actors in the phase of data acquisition: the employers or delegates who are obliged to transmit the communications, local authorities nodes and other central institutions that receive the communications and the Ministry of labour as central node, coordinating the overall system.

The CO system foresees the use of four standardized forms which can be used for a number of types of communications which in turn can specify one of several types of action.

As for the forms, the main one, the *Unificato lavoro* is used by all employers with the exception of temporary work agencies which have to use the *Unificato Somministrazione*. The form *Urgenze* is a synthetic form that must be followed by the *Unificato lavoro*, to communicate hirings that must be activated with a very short notice. Finally the *Variazioni datori di Lavoro* form is to be filled to communicate virtual transfer of workers associated to corporate changes.

All these forms are used to send the Compulsory Communications which carry out information about the following actions: Activation of a new employment relationship (contract) between an employer and a worker; Extension of a previous temporary contract; Transformation of the contract from one type to another (temporary to permanent, apprenticeship to permanent, part-time to full time); Termination of a contract before the natural deadline (not requested for contract ending at the natural deadline).

Beside the compulsory one, other types of communications are used in order to require: a Correction of a communication already transmitted, a Cancellation of a previous communication, an *Ex officio* insertion as a result of inspection activities.

Each communication enters the System, stored as a XML file, in a repository classified on the basis of the type of the form and identified through a key composed by a CO code. This *Administrative Database* is used for the purposes defined by law related to the inspections to contrast irregular work.

4.1 Data Processing

Starting from the *Administrative Database*, exclusion procedures are applied to get rid of: (a) of those communications for which a Cancellation or a Correction has been received (b) those communication sent repeatedly (c) of the non-due communication such as termination communications for contracts ending at the natural expiration date.

In the following phase CO data are processed and organized in a *Database of Transactions*. Each transaction represents a single event of activation, termination, extension or transformation of an employment relationship and it is identified by the CO code.

The final segment of the flow is composed by all the procedures, from those operated to pass from administrative units to statistical ones to those devoted to edit the classification variables, which are aimed to step up to a *Statistical Database*.

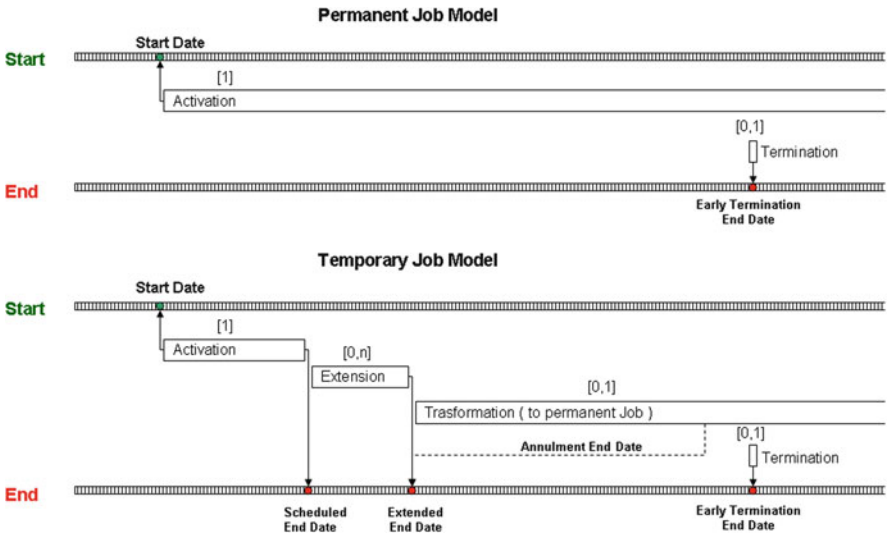

Fig. 1 Transaction sequence in permanent and temporary job reconstruction

Two intertwined classes of procedures operate in this step. A first class aims to check and edit the variables associated with the employer, the worker and the job. The modalities of each variable modality are checked against the standard dictionaries and classifications in order to identify coding errors and non-responses. It is important to stress that the procedures have to be continuously updated to keep up with the legislation changes that imply change in the forms.

In a future stage of the project it is foreseen that this step will also include an integration with other administrative or statistical sources either to edit the common variables or to add information to the system. A first example of these procedures can be the editing of the Nace code through the integration with the Business Register Asia.

The second class of procedures aims to reconstruct the main units of the statistical database, the worker, identified through the tax code and characterized by personal and biographical variables such as gender, age, educational level, nationality, address; the employer identified through tax identification code (ID) and characterized by economic activity sector, registered office, place of work or local unit; the job, defined as an employment relationship between an employer and a worker and characterized by the relationship starting date. The jobs are built by linking sequentially all the transactions referred to the same relationship: one activation, (possibly) one transformation of a fixed-term employment to permanent, (possibly) one or more extensions, and (possibly) one termination (Fig. 1).

The key that identifies the job is threefold, composed by of the ID of the worker, the ID of the employer and of the starting date of the relationship.

The variables that qualify the job are: the duration (including extensions and transformation), possibly the end date, the type of the contract (open-ended, fixed-term, seasonal), working time (full time, part-time . . .), occupation.

This is the end point of the process from which the analysis on microdata or the aggregate indicators can be produced.

5 From Transaction to Jobs: Problems, Consequences and Solutions

This last step of the process is very delicate as more records of the transaction database are recombined through record linkage procedures for which the quality of the threefold key is a prerequisite for the reconstruction of statistical information.

If one of the three key variables is affected by errors, a transaction is not matched with the correct job and the chain of events related to it is interrupted.[3] The consequence is the presence of *orphan* jobs, that is, jobs lacking a communication of activations and correspondingly the presence of jobs lacking transactions subsequent to the activation.

In principle, however, without knowing the origins, the presence of orphan jobs can be justified either on the basis that the statistical database has not received the activation communication or on the basis of non-matches to existing activations. If the first option is believed, one should impute the activation event of the orphan job using the starting date contained in the first transaction. However if this hypothesis does not correspond to reality this can result in biased estimates of the target quantities (the number of activations, terminations, the job duration, etc.). For instance, suppose that an extension of a temporary job is not merged with the corresponding activation. If the orphan job is given an activation using the information of the beginning date contained in the extension itself, the job is duplicated, the number of activation and terminations will be overestimated and the average duration of jobs is underestimated since it will include the duration foreseen in the first activation without considering that it has been moved forward though the extension. The imputation of jobs has also a negative impact on the CO data stabilization process contributing to make it longer, since it introduces a considerable lag between the date when the event occurred and that of its recording into the CO system (see the following paragraph).

The previous example should have made it clear how important is discriminating between absent transactions and non-matches.

Experimental data have allowed us to verify that the majority of cases of reconstruction of activation by different transactions run into problems due to errors in the threefold key. The test was done using the previous communication code. This code is mandatory in cases of cancellation or correction, but in some cases, although not required, is also available in the event of transformation, extension or

[3] See Baldi et al. (2011).

termination. The test has shown that the great majority of cases are due to errors in one or more of the terms of the threefold key and that the cleaning of these terms would reduce the key error of over 80 %. On the opposite considering the activation as missing would imply the duplication of a large part of the cases.

Thanks to the experience matured so far, and a procedure to reconstruct the jobs is currently under development. This will be firstly based on some edit rules whose violations will imply the provisional discard of the transaction. These can be classified into three groups: rules on the quality of the threefold key; rules regarding the inclusion of the jobs with respect to the CO system constitution date (March 1, 2008); and rules on the job end date. With respect to the first group, it is imposed that the key variables can be used only if they have passed the edit rules, before or after an imputation procedure aimed at cleaning up the threefold key. This procedure can exploit, first of all, auxiliary information present in the source itself (the recalculation of the tax code of the worker from the personal data, a further identification code of the company—the regional insurance code—an algorithm for identifying recurring error in typing the starting day of job). Furthermore residual errors might be coped with integrating external archives or adopting probabilistic algorithms.

With respect to the second group of rules, it is established that a job with a beginning date after the March 1, 2008 can be created only by an activation communication. Stemming from the legal obligation this rule avoids to create new jobs from communications other than activations with the risk of duplications. The third group of rules aims at making the end date consistent with the type of contract reported in the activation communication or with any subsequent extension or transformation, avoiding misalignments of the dates. The violation of the last two group rules can also imply inaccurate measures of the target quantities and a longer data consolidation process.

Given the complexity of the task at hand a further possibility that can be used as a provisional solution, is to use macro-like adjustments to produce estimates of parameters like the number of terminations per period (Baldi et al. 2011).

6 The Issue of Data Consolidation and Its Monitoring

A precondition to issue reliable short-term statistics is that the process of data consolidation into the statistical database is under control. For this reason the monitoring of the causes of possible delay in the transmission or in the processing is of utter importance.

Even if the employers are required to forward the activation communications within the day before the beginning of the employment relationship, a long acquisition process has been sometimes observed. The factors which contribute to make the CO data stabilization process longer can be classified into three main groups: administrative factors; data transmission problems and factors connected to processing and management of the information acquired at central level. The first group includes the *ex officio* communications (which are referred to events not declared by the employers but entered into the system only after administrative

inspections) and possible readjustment of the system as a result of regulatory changes. An example of this last factor is the long delay in the acquisition process of the domestic work communications caused by the amendment, which came into force on 15 March 2009, stating that those communications have to be transmitted via the Social security institute (Inps).

The second group of factors derives from possible problems in the data transmission from the local nodes to the central one. They may cause either blocks, and then peaks of communications, or a wrong classification of the information acquired. The main factor belonging to the third group regards problems arising in the matching of the events related to the same job due to errors in the threefold key (see paragraph 5).

An empirical study has been conducted observing the time lags needed for completing the data acquisition process. The effects of the two administrative factors above mentioned have been analysed through an analysis of the activation communication time series broken down by region and economic activity sector. The comparison of the percentage shares of total activation communications related to events occurred in the month of April 2009 with the aggregate that excludes *ex officio*, and domestic work communications has showed the negative effect of these two factors in the data stabilization process.

Furthermore, in order to compare the acquisition process dynamics referred to activations occurred in different months, we have also simulated the process assuming that it ends within a year. The results of this further analysis support the evidence of a quite gradual data stabilization process.

Although CO data should follow a unique, centrally defined, set of standards, not all the implemented applications are totally compliant with them. A part of data is acquired through massive sending, aimed at facilitating the simultaneous delivery of a large number of communications. The heterogeneity of the data capturing tools can also affect the data consolidation process.[4]

7 Conclusions and Future Developments

In this chapter we illustrate the main steps of the CO data processing, starting from the administrative declarations and ending with the construction of the complex statistical unit "job", discussing which problems can arise in the procedure flow and their statistical impact. While the structure of the entire process is already in place, some procedures are going to be refined and others set up. Four main areas of work are under development or will be started in the near future: the check and editing procedure of the threefold key, the design of consistency rules for the non-key variables, the integration of external sources either to edit variables or to add information to the system, the design of a set of indicators to monitor the quality of the process and to signal anomalies when they occur.

[4]Consolidation aspects are described in Di Bella et al. (2011).

References

Di Bella, G., De Blasio, G., Callori, M., Lucarelli, A.: A new administrative source on employment flows: aspects of the data consolidation process for statistical use. SIS-VSP Workshop on Enhancement and Social Responsibility of Official Statistics Rome, 28–29 April 2011

Baldi, C., De Blasio, G., Manieri, M., Mondauto, L., Rizzi, R.: The statistical units of the compulsory communications and the construction of jobs. Workshop-vsp2011 (2011)

Anastasia, B., et al.: Guida all'uso delle comunicazioni obbligatorie nel monitoraggio del mercato del lavoro. I Tartufi n.36, Agenzia Veneto Lavoro (2010) (in Italian)

Italia Lavoro: L'analisi delle comunicazioni obbligatorie. Nota metodologica (2010) (in Italian)

Ministero del Lavoro e delle Politiche Sociali: Comunicazioni Obbligatorie. Modelli e Regole (2010) (in Italian)

Credit Stress Testing from a Portfolio Perspective

Tiziano Bellini

Abstract

Stress testing has become an important topic since the development of the risk management practice and the enforcement of banking international supervisory requirements. The regulatory perspective is mainly focused on stressing risk parameters such as default probability (*Pd*), loss given default (*Lgd*) and so on. Thus, focusing on the lack of a well-consolidated model to link macroeconomic factors and internally estimated risk parameters, we propose a robust model to be exploited for stress testing purposes. Moreover, when extreme scenarios take place, the standard Normal distribution underlying the (Basel II) internal rating-based (IRB) formula does not necessarily hold. For this reason, we analyze the distribution of risk factors concentrating on macroeconomic Italian data from 1990 to 2009. Finally, relying on two stylized banking credit portfolios, we compute the Value at Risk (*VaR*) and the Expected Shortfall (*ES*) for alternative macroeconomic distributions.

1 Introduction

In the last few years, stress testing has become a crucial theme in the financial literature. This topic has been explored focusing essentially on the impact of macroeconomic shocks on the financial system as a whole. Bunn et al. (2005), Sorge and Virolainen (2006) and other authors emphasize the link between macroeconomic factors and systematic risks. More recently Drehmann et al. (2010)

T. Bellini (✉)
Università di Parma, Via Kennedy 6, 43100 Parma, Italy
e-mail: tiziano.bellini@unipr.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_20,

investigate the stress testing process from a different point of view. They emphasize the role of banking risk management and the impact of stress testing in assessing the effectiveness in managing risks.

Following this latter perspective, the European Banking Authority (EBA) defined a stress testing framework to be exploited to assess banking capital when extreme events take place. EBA (2011) stress testing process relies on the shock of risk parameters (i.e., *Pd*, *Lgd* and so on). From a practical point of view, however, the banking authority does not specify how to stress the abovementioned parameters. Remarking that internal ratings are usually estimated focusing mainly on debtor-specific information, the need to link internal risk parameters and macroeconomic variables arises. One important contribution of our research is to model this relationship. We apply our analysis to the index of default rates computed by the Bank of Italy and to a set of Italian macroeconomic variables from 1990 to 2009. Aiming to obtain robust estimates we exploit the forward search (Atkinson et al. 2004). This technique is based on the idea of building up an initial subset of few units and add one additional unit at each step of the search. The increasing subset is made up by those units that are the closest according to a predefined measure.

Pursuing the goal to investigate extreme scenarios, it is evident that the standard Normal distribution underlying Basel II IRB formula does not necessarily hold. According to our knowledge, there is no literature devoted to investigate the distribution of risk factors to be exploited for stress testing purposes. Thus, emphasizing the need to fit a distribution to real data, we exploit the forward search cluster analysis to identify groups of units on which to fit such distributions. Once obtained the abovementioned estimates, we generate macroeconomic scenarios. Finally, we apply these scenarios to two stylized credit portfolios computing the *VaR* and the *ES* for alternative macroeconomic distributions.

The paper is organized as follows. In Sect. 2 we describe our credit stress testing framework. In Sect. 3, exploiting the forward search approach, we analyze a set of Italian quarterly time series from 1990 to 2009. In Sect. 4 we apply our analysis to two stylized banking credit portfolios. Section 5 contains concluding remarks.

2 Credit Stress Testing Framework

According to the growing interest for banking macroeconomic stress testing, the following two issues must be addressed:

- Analysis of the link between credit risk parameters and macroeconomic variables.
- Analysis of the multivariate distribution of macroeconomic variables.

In what follows, we describe how to face these topics in a credit portfolio framework. In particular, we consider n banking debtors over a fixed time horizon (e.g., 1 year). We denote ξ_i the default indicator of debtor i, $(i = 1, \ldots, n)$. In the case where $\xi_i = 1$, debtor i defaults within the time horizon, otherwise $\xi_i = 0$. The net loss associated with the default of the debtor i is $c_i \geq 0$. In some models

c_i is assumed to be a random variable, but here, without loss of generality, we assume c_i to be a constant. The portfolio loss over the specified time horizon can be represented as follows:

$$L = \sum_{i=1}^{n} \xi_i c_i . \tag{1}$$

We can define the *VaR* at level $(1 - \alpha)$ as the smallest loss, l, such that the probability that the loss L exceeds l is no larger than α

$$VaR_{(1-\alpha)} = \inf_{l} P(L > l) \leq \alpha. \tag{2}$$

It is evident, however, that *VaR* suffers from two severe deficiencies if considered as a measure of downside risk. In fact, it is not sub-additive and it is insensitive to the size of loss beyond the pre-specified threshold level. For these reasons, in addition to *VaR*, we focus on *ES* which is the expected loss exceeding the *VaR*

$$ES_{(1-\alpha)} = E[L|L > VaR_{(1-\alpha)}] = \frac{1}{\alpha} \int_{r>1-\alpha}^{1} VaR(r)dr. \tag{3}$$

ES is sub-additive and it provides information about the amount of loss exceeding *VaR*. Then, portfolios with a low *ES* should also have a low *VaR*. In addition, under general conditions, *ES* is a convex function and it is a coherent measure of risk as well (Acerbi and Tasche 2002).

The marginal default probability $Pd_i = P(\xi_i = 1)$ constitutes the crucial element of the portfolio loss. It becomes the key issue of the credit stress testing framework described in the next section where we model the link between *Pd* and macroeconomic factors.

2.1 Credit Risk Parameters and Macroeconomic Factors

It is usual for banks to estimate internal ratings applying statistical models (e.g., logit, probit) on debtor-specific information. Thus, in order to carry out a macroeconomic stress test, it is necessary to figure out how to link internally estimated *Pd* (based on microeconomic information) and macroeconomic variables. Highlighting that there is not a well-consolidated approach, we develop a framework based on (Wilson 1997) logit model. Pointing out that our framework can be extended to other models (e.g., probit), we assume that

$$Pd_{i,s} = \frac{1}{1 + e^{-\vartheta_{i,s}}}, \tag{4}$$

where $\vartheta_{i,s}$ is a linear function of both microeconomic factors and the index Z_s, which measures the health of sector s, $(s = 1, \ldots, q)$, on which i operates. Aiming to catch the relationship between micro and macroeconomic factors, we exploit the following equations:

$$\vartheta_{i,s} = \gamma_{i,s,0} + \gamma_{i,s,1}\eta_{s,1} + \ldots + \gamma_{i,s,g}\eta_{s,g} + \gamma_{i,s,z}Z_s + \epsilon_{i,s}, \tag{5}$$

$$Z_s = \beta_{s,0} + \beta_{s,1}X_1 + \ldots + \beta_{s,p}X_p + \epsilon_s, \tag{6}$$

where $\eta_{s,v}$, $(v = 1, \ldots, g)$ are the micro, while X_k, $(k = 1, \ldots, p)$ are the macroeconomic factors on which the regression is carried out. In this latter regression we consider T observations $(t = 1, \ldots, T)$.

Considering the interaction between macroeconomic factors and default probabilities, we capture the effect of macroeconomic shocks as follows:

$$Pd_{i,s,\Delta} = \frac{1}{1 + e^{-(\vartheta_{i,s} + \Delta Z_s)}}, \tag{7}$$

where Δ stands for the shock and, in order to compute ΔZ_s, we exploit the difference between the inverse logit of $Z_{s,\Delta}$ (random generation) and z_T (historical Z_s index realization at time T). $Z_{s,\Delta}$ is computed applying a shock on each macroeconomic variable X_k of Eq. (6) as follows:

$$Z_{s,\Delta} = \hat{\beta}_{s,0} + \hat{\beta}_{s,1}X_{1,\Delta} + \ldots + \hat{\beta}_{s,p}X_{p,\Delta}, \tag{8}$$

where $X_{k,\Delta}$ denotes the shocked macroeconomic factor k, $(k = 1, \ldots, p)$.

Pursuing stress testing purposes, we need to analyze extreme and negative settings where portfolio inefficiencies could produce the highest losses. In the next section we describe how to derive the distribution of macroeconomic factors and obtain the portfolio loss distribution.

2.2 Risk Factor Distribution and the Stress Testing Framework

A stress testing process is based on extreme event analysis. One of the main contributions of our research is to integrate this analysis into a robust statistical framework. We identify macroeconomic worst events applying the forward search cluster analysis. Thus, we fit probability distributions on units belonging to clusters. Once fitted these distributions, we generate scenarios to simulate the portfolio loss distribution. More in detail, for each simulated scenario we compute shocked *Pd* and, applying Eq. (1), we compute the portfolio loss. Repeating this procedure we obtain the portfolio loss distribution. We can summarize our framework as follows:

- We consider as inputs the probabilities of default $Pd_{i,s}$, the regression parameters of Eq. (5) and the net loss c_i associated with debtor i.

- For each sector s, we estimate regression parameters of Eq. (6), considering p macroeconomic variables.
- We carry out the forward search cluster analysis on the historical matrix $\mathbf{X}$ obtaining v clusters.
- We fit the distribution of each cluster $\mathbf{X}_v$, thus we simulate the shocked macroeconomic factors $X_{k,\Delta}$.
- Starting from $X_{k,\Delta}$ and exploiting Eq. (8), we obtain the shocked *Pd* of Eq. (7). Thus, applying Monte Carlo simulations to generate the indicator function ξ of Eq. (1), we simulate the portfolio loss. From this loss distribution we derive both the *VaR* and the *ES*.

In the next section we carry out the macroeconomic analysis which is necessary to estimate robust parameters and identify clusters to be exploited in our stress testing framework.

3 Robust Macroeconomic Analysis

Our macroeconomic analysis relies on Bank of Italy and ISTAT quarterly datasets[1] from 1990 to 2009. Entering into details, we consider as Z_s, which measures the health of sector s, the index of default rates computed by the Bank of Italy. We concentrate on $q = 3$ sectors: consumers, household firms, and non-financial companies. They constitute significant axes of analysis for a credit portfolio and, considering that our goal is to show the mechanics of our framework, we do not go in depth with further regional or field analysis. The set of macroeconomic variable $\mathbf{X}$ is constituted by quarterly observations of: gross domestic product (GDP), total free on board exports (EXP), total free on board imports (IMP), gross housing investments (INV), internal consumption (CONS), and added value for food, beverages and tobacco (VAL). This choice is due, on the one hand, to the availability of these data for the entire period and, on the other, because we are mainly concerned on the analysis of real factors avoiding to consider financial time series for our study. In the next section we analyze default rates while the following section is dedicated to fit macroeconomic multivariate distributions to be exploited in our framework.

3.1 Default Rates Regression Analysis

The first task in our analysis is to estimate regression parameters $\hat{\beta}_{s,k}$ ($s = 1, \ldots, 3$; $k = 1, \ldots, 6$) of Eq. (6) where, as above described, Z_s is the index of default rates computed by the Bank of Italy for consumers, household firms, and non-financial companies, while $\mathbf{X}$ is the $T \times p$ matrix of macroeconomic variables previously

[1] Data are available on the following two websites: http://www.bancaditalia.it and http://www.con.istat.it.

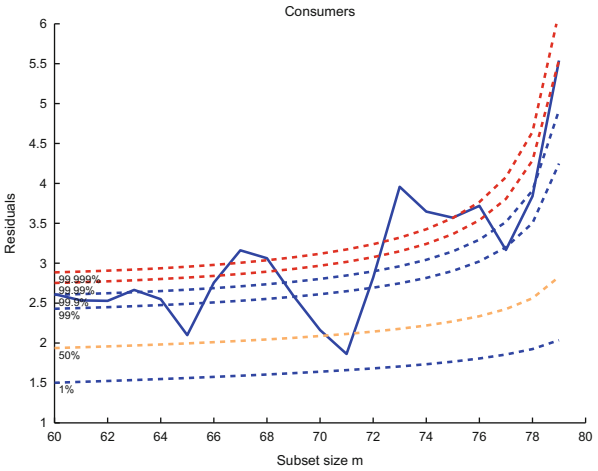

Fig. 1 Regression analysis of Bank of Italy default rates for consumers. $e_{\min}(m)$ *solid-curve* is compared to *dotted-curves* envelopes at different confidence levels. It is shown that in the last part of the search there are observations well beyond the extreme upper thresholds

described. Highlighting the need to estimate these parameters in a robust way, we exploit the forward search which is carried out through the following main activities: choice of the initial subset, addition of observations during the search and monitoring statistics step by step along the search (Atkinson et al. 2004). In regression, as highlighted in Fig. 1, we concentrate on the monitoring of residuals. In particular we focus on the progress of $e_{\min}(m)$, the minimum residual for units not belonging to the subset S_m, as the subset size m increases. Large values of the residuals among units not in the subset indicate the presence of outliers.

We start our data analysis concentrating on the consumer default rates. In Fig. 1, the solid-line curve, which represents $e_{\min}(m)$, is compared to dotted-line envelopes which, according to Riani et al. (2009), constitute theoretical boundaries, at different confidence levels, for the inference on outliers. As it is evident from Fig. 1, at step 67 the solid-line curve $e_{\min}(m)$ crosses envelopes at all confidence levels (even at 99.999 %). Then we superimpose envelopes in order to verify whether outliers are detected. In this analysis 13 observations are identified as outliers. An immediate question arises: *Is there a useful transformation for our data?*

We answer to this question applying the forward search parametric family transformation to default rates for: consumers, household firms, and non-financial companies. Concentrating on the commonly used set of power transformation parameters $\lambda \in \{-1, -0.5, 0, 0.5, 1\}$, the most suitable transformation for all time series is $\lambda = -0.5$. We perform the regression analysis of Eq. (6) on the transformed data. In the case where we do not apply any transformation, outliers are detected in both consumers and non-financial companies default rate time series. By contrast, in the case where we apply the transformation with $\lambda = -0.5$, no outliers are detected.

The regression on transformed data leads to high fitting levels showing R^2 values of 88, 81, and 76 %, respectively, for consumers, household firms, and non-financial companies.

3.2 Cluster Analysis Through the Forward Search

In order to identify the macroeconomic distribution to simulate portfolio losses, it is useful to investigate macroeconomic time series. In particular, we investigate whether statistical units can be grouped into cluster on which to fit a multivariate distribution. Furthermore, in the case where extreme events can be isolated, we can fit a specific distribution on them applying such an extreme distribution for stress testing purposes. According to this perspective, the forward search cluster analysis becomes one of the key elements of our analysis. This technique relies on the analysis of the minimum Mahalanobis distance $d_t(m)$ among observations not in the subset

$$d_{\min}(m) = \min d_t(m), \qquad t \notin S_m. \tag{9}$$

If this observation is an outlier relative to the other m observations, this distance will be large compared to the maximum Mahalanobis distance of observations in the subset. All other observations not in the subset will, by definition, have distances greater than $d_{\min}(m)$ and will therefore be outliers.

We execute the abovementioned procedure exploiting a trimming multi-step approach. As described in Atkinson et al. (2004), a preliminary study is carried out followed by an explanatory and a confirmatory analysis to which corresponds the final allocation of units to clusters.

We apply the forward search cluster analysis on the $T \times p$ matrix $\mathbf{X}$ of the following macroeconomic variables: GDP, total free on board exports (EXP), total free on board imports (IMP), gross housing investments (INV), internal consumption (CONS), and added value for food, beverages and tobacco (VAL). As it is evident from Fig. 2, the distributions of these variables, apart from VAL, are very far from the Normal ones, then we cannot exploit Box–Cox transformations as in the previous section.

Through our analysis, we obtain three clusters and a set of outliers. It is evident from Fig. 2 that clusters can be linked to alternatives states of the economy which we name: *high, middle*, and *low*. At the same time, it is useful to notice that atypical units behave very differently from other units at least from one-or two-dimensional analysis.

We fit the macroeconomic distribution on each of these clusters. According to Eq. (7), these parameters will be used to generate macroeconomic scenarios to shock *Pd*. Once obtained shocked *Pd* the portfolio loss distribution will be derived.

In the next section we exploit the above-described analysis to investigate the portfolio loss distribution for two stylized credit portfolios.

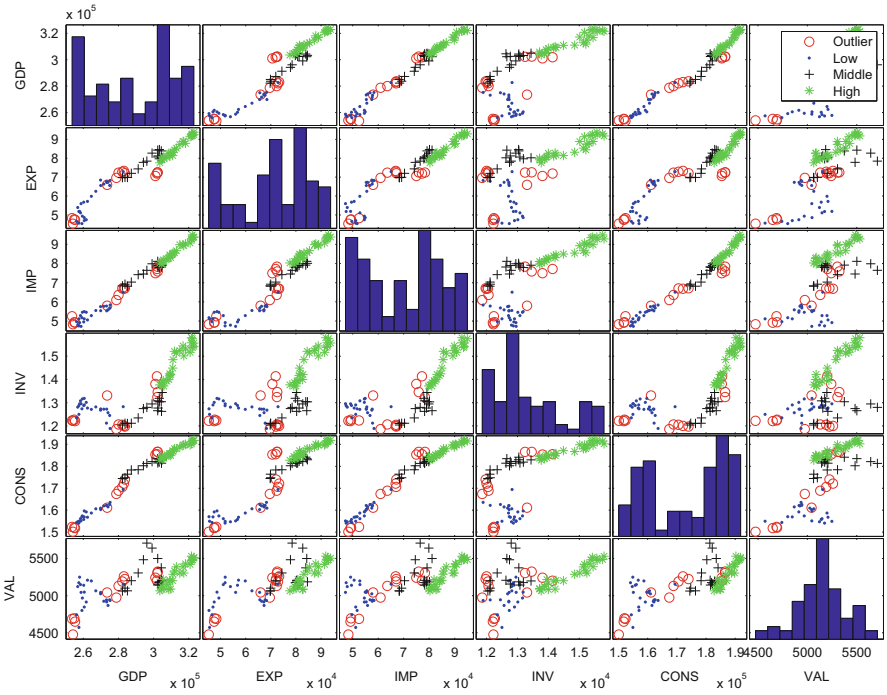

Fig. 2 Scatterplot matrix and clustering of macroeconomic variables. The final allocation of the forward search cluster analysis shows three main clusters and a set of atypical observations. These three clusters roughly correspond to the states of the economy which we define: *high, middle*, and *low*

4 Portfolio Stress Testing

In what follows we concentrate on two stylized banking credit portfolios with $n = 105$ debtors (each debtor has an exposure of 100) belonging to the following sectors: consumers, household firms, and non-financial companies. As highlighted in previous sections, we assume as given the probabilities of default $Pd_{i,s}$, the regression parameters of Eq. (5) and the net loss c_i of each debtor. In the first portfolio all debtors have the same $Pd_{i,s}$ and c_i and they are homogeneously distributed among the three sectors. In the second portfolio we relax these assumptions allowing debtors to have different risk parameters as well as non-homogeneous distributions among sectors.

We generate the default indicator function ξ exploiting Monte Carlo simulations. Figure 3 shows the loss distributions for alternative macroeconomic settings. Loss distributions named: *high, middle*, and *low* are identified in panels (a), (b), and (c). They are obtained simulating macroeconomic factors as multivariate Normal random variables. The mean vectors as well as covariance matrices are estimated on

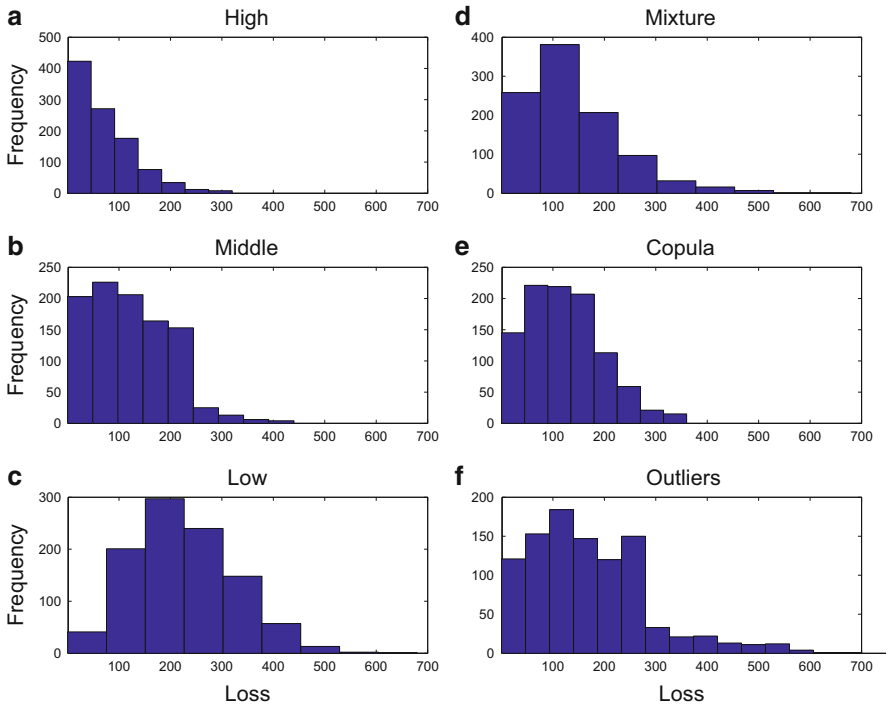

Fig. 3 Loss distribution for the first stylized credit homogeneous portfolio. The generation of macroeconomic factors from alternative parameter sets leads to different loss distributions. The most extreme losses are associated with multivariate Normal simulations with parameters estimated on the set of outliers

observations belonging exclusively to clusters: *high, middle*, and *low*. Panel (d) of Fig. 3 shows the loss distribution linked to the Normal mixture random simulations obtained considering weights and parameters estimated on clusters: *high, middle*, and *low*. In order to verify whether our findings are aligned with those deriving from one of the most commonly used models in the literature (Glasserman and Li 2005), in panel (e) we show the loss distribution due to the Normal copula approach. Finally, panel (f) highlights the loss distribution obtained relying on Normal random simulations due to means and covariances estimated on the set of outliers.

It is now evident from Table 1, where both portfolios are considered, that *VaR* and *ES* are ranked according to the abovementioned states of the economy.

When we generate macroeconomic factors considering *high* cluster parameter estimates, losses are lower than in the case of the *middle* ones. This latter distribution implies approximately the same losses as the Normal copula. Higher losses, on the other hand, are associated with generations from the *low* cluster as well as from the Normal mixture, based on all the abovementioned clusters. The most extreme losses are linked to simulations from the set of outliers. It is evident from Table 1 the

Table 1 *VaR* and *ES* for portfolios 1 and 2

	High	Middle	Low	Mixture	Copula	Outlier
Ptf 1 *VaR*	240	320	480	400	320	780
Ptf 1 *ES*	268	344	516	456	336	992
Ptf 2 *VaR*	342	449	718	580	420	873
Ptf 2 *ES*	379	499	802	651	448	1,131

The ranking is linked to *high, middle* and *low* scenario simulations. The most extreme losses are associated with simulations due to the set of outliers

difference between the risk calculated from this latter approach and the risk derived from the copula distribution which is based on the standard Normal assumption.

5 Concluding Remarks

We introduce a structure to link macroeconomic variables and default probabilities estimated on microeconomic information. Our modeling allows banks to directly exploit their internal rating estimates for stress testing purposes.

Starting from Italian macroeconomic time series from 1990 to 2009, we estimate robust parameters to be exploited for shocking default probabilities. We emphasize the need to identify extreme macroeconomic events to simulate stress testing scenarios. Thus, we carry out a robust cluster analysis. We obtain clusters which we identify according to the states of the economy: *high, middle*, and *low*. We identify an additional set of atypical units which we consider as extreme events. Starting from the above-described estimates, we simulate alternative portfolio loss distributions computing the value at risk and the expected shortfall for alternative credit portfolios. We highlight the importance to carry out a stress testing process relying not only on shocked risk parameters, as it is usual in the financial literature, but also to consider a stressed risk factor distribution.

References

Acerbi, C., Tasche, D.: On the coherence of expected shortfall. J. Bank. Finance **26**, 1487–1503 (2002)

Atkinson, A.C., Riani, M., Cerioli, A.: Exploring Multivariate Data with the Forward Search. Springer, New York (2004)

Bunn, P., Cunningham, A., Drehmann, M.: Stress testing as a tool for assessing systemic risk. Financ. Stability Rev. June, 116–126 (2005)

Drehmann, M., Stringa, M., Sorensen, S.: The integrated impact of credit and interest rate risk on banks: a dynamic framework and stress testing application. J. Bank. Finance **34**, 713–729 (2010)

EBA: EU-Wide Stress Test: Methodological Note. http://www.eba.europa.eu (2011)

Glasserman, P., Li, J.: Importance sampling for portfolio credit risk. Manag. Sci. **51**, 1643–1656 (2005)

Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc. Ser. B **71**, 201–221 (2009)
Sorge, M., Virolainen, K.: A comparative analysis of macro stress-testing methodologies with application to Finland. J. Financ. Stability **2**, 113–151 (2006)
Wilson, T.C.: Portfolio credit risk (I). Risk **10**(9), 111–117 (1997)

Remote Processing of Business Microdata at the Bank of Italy

Giuseppe Bruno, Leandro D'Aurizio, and Raffaele Tartaglia-Polcini

Abstract

Providing the possibility to run personalised econometric/statistical analyses on the appropriate datasets by remote processing allows greater flexibility for applied economic research. Binding confidentiality requirements is required with business surveys. The Bank of Italy's infrastructure BIRD pursues at the same time two seemingly conflicting aims: to allow researchers to explore and model business survey data, while preserving anonymity of individual information. The system is based on the Lissy platform (already adopted by the Luxembourg Income Study and other research centres). Firms' data confidentiality is safeguarded by a number of means. We show synthetic data about the current utilisation of the platform and outline some major future developments aimed to improve service levels, as well as to maintain and improve its capability to attract qualified users.

1 Introduction: From Surveys to Data Access

The Bank of Italy started regular sample surveys in 1965, when the first wave of the Survey of Household Income and Wealth took place. In 1974 the interviews for a survey of manufacturing firms were conducted for the first time by the Bank's branches. Since then, survey microdata have been used by economists at the Research Department for policy use as well as for economic research, and the target population of the business surveys has been extended over the years.

G. Bruno • L. D'Aurizio (✉) • R. Tartaglia-Polcini
Bank of Italy, Economics and Statistics Department, Rome, Italy
e-mail: giuseppe.bruno@bancaditalia.it; leandro.daurizio@bancaditalia.it; raffaele.tartagliapolcini@bancaditalia.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_21,

The survey (*Invind*, hereafter) has taken advantage of close contact with respondents and the possibility of laying out questionnaires tailored to the Bank's research needs. Although participation in the survey is not compulsory, there is a long-standing tradition of collaboration between Italian firms and the Bank for data collection. Firms are confident that data are collected for the purpose of economic analysis only and that confidentiality is guaranteed. A mutual climate of trust has been built over the years, favoured by the general opinion of the Bank of Italy's employees as public-minded and trustworthy civil servants.

Aggregate statistical tables for both surveys have been regularly published on the Bank's Annual Report. More detailed results of the household survey, with standard tabulations have been regularly published from the 1980s. Detailed results from the business surveys have been published in form of a dedicated report since 2003.

Data from the household surveys have always been made available to interested researchers, first on request and then on the Bank of Italy's website. On the contrary, business survey data had never been made available outside the Bank until 2008.

Historically, empirical research in economics has been hampered by the lack of microdata for a long time. Different solutions have been adopted since the 1990s to overcome this situation. In the Bank of Italy, the goal of access to microdata without breaching confidentiality commitments towards the sample firms has been regarded as a way to fulfil the commitment to transparency and accountability in the domain of applied economic research.

This chapter gives an account of the Bank of Italy's solution to enable external access to business survey microdata. The system BIRD (Bank of Italy Remote access to micro Data) built for the purpose at the beginning of 2008 is described and motivated. We also analyse the first 4 years of utilisation of the system by the users and highlight the future prospects.

The paper is organised as follows. Section 2 accounts for the confidentiality rules adopted in managing external microdata access at the Bank of Italy. Section 3 describes the data currently available on BIRD and their features. Section 4 illustrates how the platform has been employed by researchers so far. Section 5 concludes, outlining the planned improvements. Appendix 1 contains figures on system utilisation; technical details are presented in Appendix 2.

2 Data and Confidentiality Safeguards

Protecting confidentiality in microdata collected in surveys has two motivations: it is both required and sanctioned by the law and is expected by survey participants.

The Bank adheres to these two principles: its commitment towards firms participating in business surveys is clearly stated on the first page of the business survey questionnaires:

> Confidentiality notice – [. . .] The information provided will be used for research purposes only and will not be disseminated outside the Bank except in aggregate form. [. . .] The

> data will be processed entirely inside the Bank with procedures that ensure their security and confidentiality. [. . .][1]

Yet, access to microdata is increasingly asked for research purposes by the scholars' community. Therefore, public institutions that collect microdata have tried to reorganise their internal procedures to meet this demand.

In the past, data utilisation of Bank of Italy's business surveys was traditionally restricted to internal researchers in order to safeguard firms' data confidentiality. Starting from 2007, a project was launched to provide a wider access to these information. Two driving reasons finally motivated the institution to pursue this policy: (1) in order to broaden the scientific debate, the Bank of Italy's researchers increasingly wanted that other scholars could verify independently results based on internal business data, presented in seminars, workshop and conferences; (2) the most prestigious economics reviews started to demand that the reviewer could replicate the empirical results presented in a submitted paper.

Clearly, a trade-off between security and accessibility of databases emerges. Whether or not the availability of datasets for processing is an intrinsic threat to data confidentiality (regardless of the means adopted to protect it) is a hotly debated issue. An easy solution is anonymising microdata aimed at public release: this is obtained by eliminating identification variables and collapsing classifications. This option is quite safe if data variability is not too high and outliers are not present. In such a way, a public use file (PUF) is obtained that can safely be distributed. This is a viable solution for household data, where sensitive information like figures on wealth and income cannot be easily identified. The Bank of Italy chose this solution for the Survey of Household Income and Wealth (SHIW). In this case, the original data are authentically anonymous, since the company in charge of the interviews leaves the basic identification variables out of the database provided to the Bank. Before the final public release, individual records are further stripped of sensitive information (like those regarding morbidity) and other fine geographical classifications which might facilitate disclosure. Data can be downloaded from the web, together with the relevant metadata.

Anonymisation can be unsafe if there are many outliers in the dataset. This tends to happen with business surveys data collected on firms with varying sizes. In such a case, access to databases can be monitored by a number of means. In the traditional "data lab", the researcher has to show up in person at the physical place where data are made available: here she can login to the desired dataset while her processing is carefully scrutinised. Such a solution has been considered for years the only secure device for data access from external researchers. During the last decade, also thanks to widespread use of Internet services, the possibility of remote processing came into light (see, e.g. Rowland 2003) and seemed from the very beginning a neat improvement with respect to previous solutions. The underlying idea is to let researchers access the lab remotely via some secure device.

[1] Excerpt from the current version (effective starting March, 2012). The former version included an explicit reference, no longer needed, to the Italian privacy law.

The possibility to run regression models requires to implement legitimacy checks through suitable filters applied to the submitted programs. This solution is expensive in terms of the security system that must be set up, yet it is the most advanced, since it reconciles confidentiality and external usability of the datasets. A prototype called RADL (Remote Access Data Lab: Trewin 2003) is described in Keller-McNulty and Hunger (1998). The Bank of Italy adopted for business survey data this solution, under the acronym of BIRD (Bank of Italy's Remote access to micro Data).[2]

The first measures taken to safeguard data confidentiality of the remotely available datasets are the usual ones adopted in the Public Use Files. Together with the usual practice of eliminating direct identifiers and collapsing of classification cells, free-format text fields are not released, since they would make security breaches easier (this practice is also followed by Statistics Canada: Rowland 2003, *cit.*). A further tool to protect the confidentiality of individual data is a preventive treatment of all the quantitative variables. For each of them a cut-off upper value is defined and, for the top five big companies, the corresponding values are set equal to the cut-off plus a disturbance term preserving data variability. This solution provides valid inputs for all the analyses different from total estimations: should the researcher need them, she can ask to use the database with the original values (in such a case outputs will be examined more carefully; see also Ritchie 2005). Data-perturbation techniques (Reiter 2005) available via specialised software packages like Argus (Nordholt 1999) were discarded, since they might have produced results too much influenced by the perturbation techniques used. Our strategy is in line with that of the most advanced statistical institutes (see Hjelm 2010 and Capobianchi 2006).

On the administrative side, the Bank of Italy follows three principles to let external users access business survey data: (1) researcher's eligibility; (2) automatic and manual legitimacy checks on submitted commands, logs, outputs and submission syntax; (3) upgraded checks when disclosure risk increases. The eligibility is based on the following criteria: (a) personal identification by valid document; (b) proven affiliation to a research institution; (c) submission of a concise research project; (d) formal agreement with the Privacy law and the Deontological code (signed form). Legitimacy checks are based, for each package, on a subset of forbidden keywords which could potentially disclose individual information. These subsets are very small, in order to provide researchers with the widest possible access. The system entails a series of manual interventions and is therefore suitable to serve a limited number of motivated researchers.

[2]See Appendix 2 for a concise description of the technical features of the Lissy 8 platform, on which BIRD will be implemented from the beginning of 2012. A justification of the architecture chosen by the Bank of Italy for its own system is found in Schouten and Cigrang (2003).

3 The Available Data

3.1 Invind

The survey has been uninterruptedly carried out since the beginning of the 1970s, but data are available in BIRD only since 1984, when they started to be collected into electronic archives.

From 2002 the panel sample covers the population of firms belonging to the industrial and service sectors with 20 employees and more (the service sector does not include the financial institutions, for which the Bank of Italy collects census data). Nevertheless, the reference population represents a significant share of the total turnover, investment and number of employees of the whole population (respectively 70 e 60 % of payroll employment for non-construction industry and non-financial private services).

Each record contains information collected for a single firm in a survey year. Variations between two adjacent calendar years are computed within the same survey year, since, for the main variables, such as employment, turnover and investment levels, the firm reports values for the survey reference year, the previous one, plus a forecast for the interview year.[3] The survey adopts a one-stage stratified sample design. The sample size is determined in two stages. First, the number of size classes is identified using the Neyman method (*optimum allocation to strata*) which minimises the variance of the sample means of the main variables observed (turnover, investments and employment). Second, the number of units in each size class is divided among regions and branches of activity in proportion to the number of firms in the target population belonging to that stratum. The missing data of the main survey variables are imputed. In general, ratio estimators are used to impute data, setting the number of the firm's employees as denominator (this information is mandatory); in some cases the firm's time series data are used for the reconstruction. The percentage of imputed data is usually small. An indicator variable marks the imputed value. Monetary values for investment and turnover are made available at constant prices, referred to the most recent year. Deflators are obtained by averaging individual values collected within the survey.

3.2 Sondtel

From 1993 a business outlook survey (*Sondtel* from now on) has also been carried out, basically on the same sample as *Invind*.[4]The interviews take place between September and October each year. Forecasts on the firm's specific activities are

[3]If we indicate with $t-1$ and t the previous year and the survey year, $t+1$ refers to the year for which a forecast is available. Interviews are carried out in the first months of year $t+1$.

[4]*Sondtel* sample was a subset of the *Invind* sample until 2000.

collected in qualitative form during a telephone interview lasting on average 15–20 min.

Exact matching between the two surveys makes possible many analyses of the association between plans/expectation (collected in *Sondtel* in categorical form) and corresponding *Invind* realisations that have been previously categorised. Data are available both in separate archives for every year and in a historical archive containing only the variables with a long history.

3.3 Common Features

The same firm within the dataset is univocally identified by a unique identification key (called IDENT, automatically generated and unrelated with any firm characteristics). The firm's structure can change over the years through mergers and acquisitions: however, an unchanged IDENT tells that the firm has fundamentally remained the same. An expansion weight, summing up to the population's number of firms is also provided. The weight takes into account both total missing responses and post-stratification adjustments (Kalton and Flores-Cervantes 2003). Weights are not calibrated to reproduce totals for other variables (such as number of employees) because too much variability would have been added to their distribution (Deville et al. 1993). Strata are combinations of sectors of economic activities and size classes, whereas post-strata are formed by combinations of macro-areas and very aggregate classifications of economic activities and size. The weighting procedure takes place in two steps. In the first, the combinations of branch of activities and size classes are used as strata. Post-stratification reconstructs only a reduced number of combinations of geographical areas, size classes and economic sectors. An extensive documentation is available on the Bank of Italy's website[5] with all the information needed for BIRD users.[6] The survey questionnaires are also available. The documentation is updated every year.

4 The System Utilisation[7]

The system became operational in February 2008. Thirty-two researchers have so far applied to submit jobs using the BIRD system. Fifteen users actually submitted jobs; the total number of submissions is 1,223. Figure 1 shows the number of submissions, by quarter. The pattern is seasonal: submissions concentrate in the first 6 months of

[5] See Bank of Italy (2012a, b) (various years) for the evolution of the survey and the main results.

[6] http://www.bancaditalia.it/statistiche/bird.

[7] Based on data available on 20 June 2012.

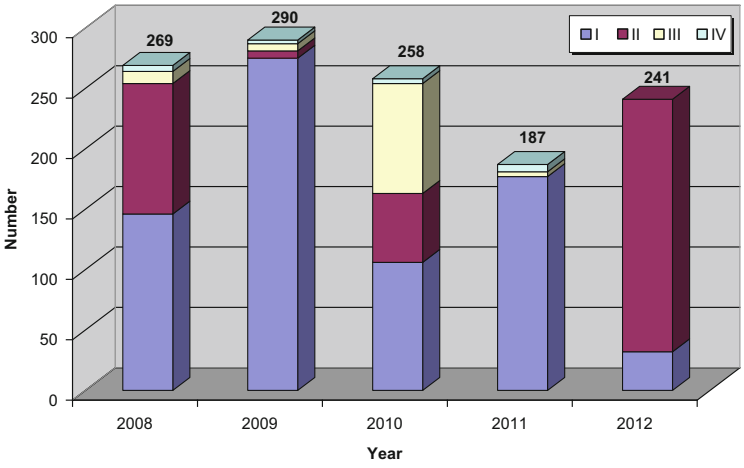

Fig. 1 Number of program submissions, by quarter

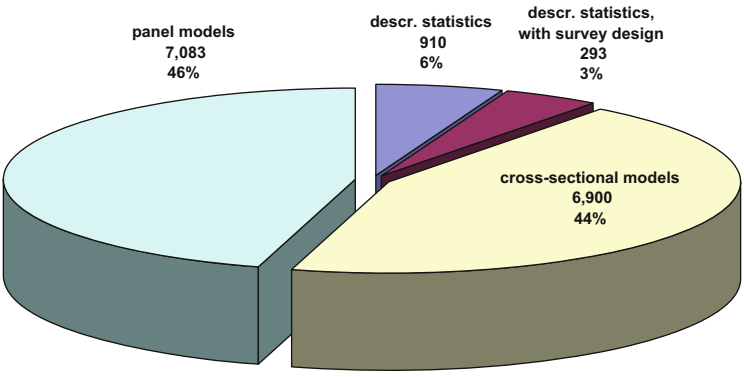

Fig. 2 Types of analyses performed

the year. This is probably correlated with the fact that the news of the release of the updated datasets is provided on the Bank of Italy's website at the beginning of each year.

All the programs were submitted in Stata. A possible taxonomy of the ways of exploitation of the available data include four broad types: (1) simple descriptive statistics (frequency distributions, means, totals, percentiles, etc.) computed without taking into account the survey design; (2) same as above, but taking into account the survey design; (3) cross-sectional regression models; (4) panel regression models. A single program could encompass more than one type of data utilisations. Figure 2 provides the corresponding frequencies for BIRD users.

We can conclude that this small group of users is mainly interested in econometric modelling, often with quite sophisticated techniques such as Generalised Method

of Moments (GMM). Panel estimation methods were chosen more often than cross-sectional ones. Descriptive statistics are generally found in the exploratory stage at the beginning of the program and are ancillary to the following model specification steps. Only in a minority of cases were means, ratios and frequencies derived by taking into account the sample design.

5 Conclusion: Providing New Services to the Users

In the last 20 years, economists and policymakers have enriched their statistical analyses through wider microdata access (Abowd and Lane 2003) as they became increasingly less satisfied with the standard statistical outputs released in official publications. The main driver of this pressure has been the desire to exploit detailed information for some sectors of the economy not available elsewhere. Moreover, the explanatory power of micro-econometric techniques started to be widely recognised.

Public bodies, however, face two constraints when they grant wider access to data. Data confidentiality should be guaranteed, because either it is mandated by law, or it is a commitment towards the respondents, or both. The need to safeguard confidentiality is particularly felt in business surveys. Firms' representatives disclose details of their strategies under the assumption that these information will not be revealed. The second constraint concerns the resources to devote to related services: these are not negligible since a complex organisational structure must be maintained.

The remote access system for Bank of Italy's business survey microdata was built in order to pursue the accountability principle, engrained in central banks' behaviour also in the domain of data dissemination, while maintaining the confidentiality commitment towards the respondents. On one hand, the availability of new data enables production of new analyses; on the other hand, results obtained by internal researchers are thus verifiable and replicable by other scholars.

We have recently begun to implement some new features in order to improve the attractiveness of the platform. They are tailored to users willing to apply complex econometric models to business survey data, with the aim of writing high quality papers in applied economics. Starting from the end of 2011, we have provided customised datasets on request, formed by joining the standard information available on the BIRD platform with additional data provided by the user. The user must first exhibit the rights to exploit the external data. The new variables will be treated according to the standard security rules, in order to preserve for the merged dataset the original level of confidentiality. Two users have exploited this possibility so far.

Since 2013, we make the R statistical package available on the platform. R is an open source package, available free of charge, possibly to become a de facto standard for developing statistical software: it has many features that could be appealing for the typical user of the BIRD platform.

Appendix 1: System Utilisation, 2008q1–2012q2[8]

See Figs. 1 and 2.

Appendix 2: The IT Architecture of Lissy 8: User Interface and Security Rules

The technical infrastructure is based on a standard Intel machine running the Windows operating system. The software application offering the remote processing capability (henceforth Lissy: Cigrang and Coder 2003) is based on a standard Java front-end application relying on a relational Data Base Management System back-end which stores all information pertaining the users, the jobs and the package configurations.

The wish to provide the users with the maximum flexibility of a statistical package has led us to the employment of the following two communications means: (1) simple text mail for communicating from and to the external users; (2) a web interface. All incoming jobs are logged as well as all system performance parameters, providing very precise statistics about usage and performance of the system.

Lissy 8 system is based on: (1) the PostOffice (PO) Server; (2) the Batch Machines (BM) (Fig. 3). The PO Server is the hub of the whole application. The BMs host independently the various package engines. By using suitable parameters, it is possible to apply differently priority rules to the jobs according to different access levels and computational needs. The data server is the repository for all of the available datasets. Separate directories are maintained for each file formats required by the available statistical packages. There is no direct link between the data and the mail server.

Lissy 8 implementation at the Bank of Italy includes SAS (until 2012), Stata and R packages. The system's users do not access the package engines directly but only through the mediation of the PO server and the BMs.

Cautionary controls are based on the technical features of the communication. First, TCP/IP traffic is allowed only on predefined ports. Second, the incoming e-mail must be plain text and carry no attachments. A final confidentiality preserving check is based on a set of forbidden words or sequences. If no problems are revealed the job is executed, otherwise execution is denied and the user notified. It is also possible to create a list of "borderline" commands: in such cases, job execution is put under manual control.

The Web-Tabulator is a Java software component designed to satisfy the informational needs of those interested in obtaining bi-dimensional tables of frequencies or other descriptive statistics. The user is presented with a menu of different datasets and an easy to use drag-and-drop procedure to build a table with two categorical

[8]Based on data available on 20 June 2012.

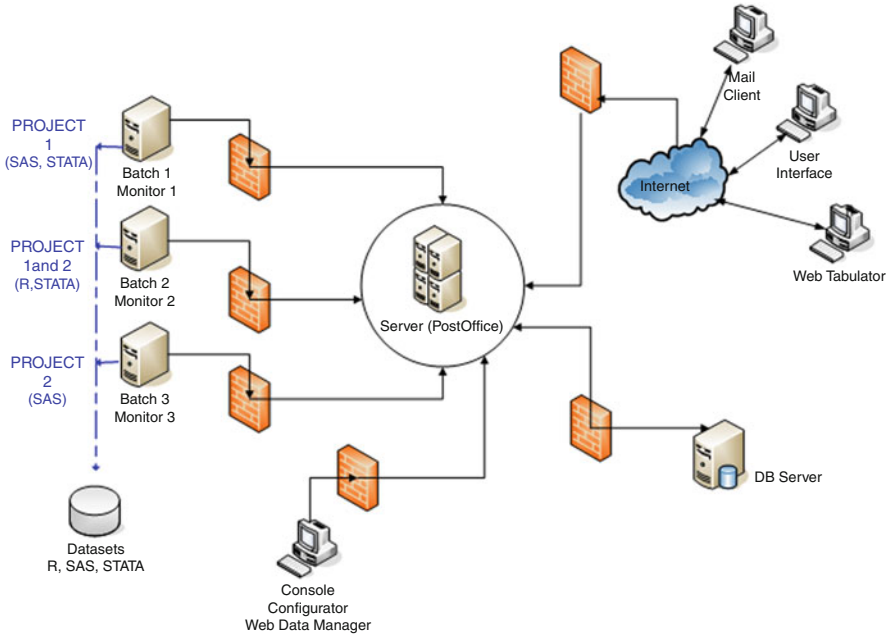

Fig. 3 The Lissy 8 platform

variables and some continuous variables. The tables produced can be exported in formats like CSV etc.

Datasets added to the Web-Tabulator are organised in dataset pools. The fundamental idea of dataset pools is to organise the data in order to collect datasets that are comparable and compatible. As the Web-Tabulator allows to perform statistics and displays results for multiple datasets at once, they must have common continuous variables to perform the statistics on as well as common discrete variables in order to sort and display the calculated results in a table. For the sake of confidentiality, for each of the dataset pool added, the minimum number of records that must be reached to display a result in a cell of the final table has to be defined.

References

Abowd, J.M., Lane J.I.: Synthetic Data and Confidentiality Protection. Technical Paper No. TP-2003-10. US Census Bureau, LEHD Program (2003)

Bank of Italy: Supplements to the Statistical Bulletin – Sample Surveys – Survey of Industrial and Service Firms. http://www.bancaditalia.it (2012a) (several years). Accessed 12 May 2014

Bank of Italy: Supplements to the Statistical Bulletin – Sample Surveys – Short term Outlook Survey of Industrial and Service Firms. http://www.bancaditalia.it (2012b) (several years). Accessed 12 May 2014

Capobianchi, A.: Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa. Documenti Istat (2006)

Cigrang, M., Coder, J.: Lissy remote access system. Paper presented at the Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg, 7–9 April 2003

Deville, J.C., Särndal, C.E., Sautory, O.: Generalized raking procedures in survey sampling. J. Am. Stat. Assoc. **88**(423), 1013–1020 (1993) (Theory and Methods)

Hjelm, C.G.: Mona-microdata on-line access at statistics Sweden. Mimeo, Statistics Sweden (2010)

Kalton, G., Flores-Cervantes, I.: Weighting methods. J. Off. Stat. **19**(2), 81–97 (2003)

Keller-McNulty, S., Hunger, E.A.: A database system prototype for remote access to information based on confidential data. J. Off. Stat. **14**(4), 347–360 (1998)

Nordholt, E.S.: Statistical disclosure control of the statistics Netherlands employment and earnings data. Paper presented at the Joint ECE/Eurostat work session on statistical data confidentiality, Thessaloniki, 8–10 March 1999

Reiter, J.: Estimating risks of identification disclosure in microdata. J. Am. Stat. Assoc. **100**(472), 1103–1112 (2005)

Ritchie, F.: Access to business microdata in the UK: dealing with the irreducible risks. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, 9–11 November 2005

Rowland, S.: An examination of monitored, remote microdata access systems. Paper presented at the national academy of sciences workshop on access to research data: assessing risks and opportunities, Washington, DC, 16–17 October 2003

Schouten, B., Cigrang, M.: Remote access systems for statistical analysis of microdata. Stat. Comput. **13**, 381–389 (2003)

Trewin, D.J.: Access to microdata: issues, organisation and approaches. Paper presented at the conference of European statisticians, Geneva, 10–12 June 2003

A Distributional Approach for Measuring Wage Discrimination and Occupational Discrimination Separately

R. Giaimo and G.L. Lo Magno

Abstract

The well-known Blinder–Oaxaca [Blinder, J. Hum. Resour. **8**(4), 436–455 (1973); Oaxaca, Int. Econ. Rev. **14**(3), 693–709 (1973)] decomposition divides the wage differential between men and women into a part, which can be explained by differences in individual characteristics, and another part, which is usually interpreted as discrimination. This decomposition neglects any distributional issues in evaluating discrimination, thus permitting undesirable compensation between positively and negatively discriminated women. Jenkins [J. Econ. **61**(1), 81–102 (1994)] has criticized this aspect, instead, preferring a distributional approach, where the entire distribution of experienced discrimination is evaluated. Following Jenkins [J. Econ. **61**(1), 81–102 (1994)], Del Río et al. [J. Econ. Inequal. **9**(1), 57–86 (2011)] use a distributional approach, adapting the Foster–Greer–Thorbecke [Econometrica **52**(3), 761–766 (1984)] class of poverty indices to the study of discrimination.

Studies adopting this approach merit little attention as regards the issue of the separate measuring of wage discrimination and occupational discrimination. Alternatively, we have used the Foster–Greer–Thorbecke indices for measuring wage discrimination and occupational discrimination separately. Similar to the technique employed in the Brown–Moon–Zoloth decomposition [J. Hum. Resour. **15**(1), 3–28 (1980)], we have employed a multinomial model

R. Giaimo
Department of Agricultural and Forestry Sciences - Faculty of Agriculture - University of Palermo, Viale delle Scienze, Building 4, 90128 Palermo
e-mail: rosa.giaimo@unipa.it

G.L. Lo Magno (✉)
Department of Economics, Business and Statistics - Faculty of Economics - University of Palermo, Viale delle Scienze, Building 13, 90128 Palermo
e-mail: lomagno.gl@virgilio.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_22,

for estimating the theoretical distribution of women in occupations, which would result in the absence of occupational discrimination.

Introduction

The standard approach to measuring wage discrimination is the Blinder–Oaxaca decomposition (B–O) (Blinder 1973; Oaxaca 1973), in which the hourly wage differential between men and women is decomposed as follows:

$$\ln \overline{W}_{\mathrm{M}} - \ln \overline{W}_{\mathrm{F}} = \left(\overline{Z}_{\mathrm{M}} - \overline{Z}_{\mathrm{F}}\right) \widehat{\beta}_{\mathrm{F}} + \overline{Z}_{\mathrm{M}} \left(\widehat{\beta}_{\mathrm{M}} - \widehat{\beta}_{\mathrm{F}}\right) \tag{1}$$

where $\ln \overline{W}_{\mathrm{M}}$ and $\ln \overline{W}_{\mathrm{F}}$ are the means of the logarithms of observed hourly wage of men and women respectively, $\overline{Z}_{\mathrm{M}}$ and $\overline{Z}_{\mathrm{F}}$ are mean vectors (calculated for the observed sample) of individual characteristics, which are believed to affect wage, and $\widehat{\beta}_{\mathrm{M}}$ and $\widehat{\beta}_{\mathrm{F}}$ are OLS estimates, which are obtained by regressing, separately by sex, logarithm of hourly wage on those characteristics. The first part of the decomposition represents the wage differential explained by differences in individual characteristics, while the second is usually interpreted as discrimination. In the decomposition presented above, the differences in remuneration rates given by OLS estimates for regression coefficients are weighted by $\overline{Z}_{\mathrm{M}}$, while the differences in average endowments are weighted by $\widehat{\beta}_{\mathrm{F}}$. Other analogue decompositions, using different weightings, are provided by Reimers (1983), Cotton (1988), Neumark (1988) and Oaxaca and Ransom (1994).

Jenkins (1994) has criticized this standard approach because it does not adequately take into account the distribution of wage discrimination experienced by each woman. Indeed, it can be shown that the evaluating of wage discrimination, performed with the Blinder–Oaxaca decomposition, can lead to the conclusion of an absence of discrimination when positively discriminated women are compensated by negatively discriminated women, even when there is no conceptual doubt that discrimination is present. Moreover Jenkins (1994) has underlined a common aspect of poverty and discrimination: both can be viewed as a form of deprivation. Regarding poverty analysis, deprivation derives from a poverty line; in the case of discrimination, deprivation results from the wage which women would receive if no discrimination penalized them. In order to focus on distributional issues of discrimination, the distributional approach employs a two-step framework of poverty analysis: (1) defining a measure of individual discrimination for each woman; and (2) defining an index to summarize the entire distribution of the individual female discrimination. This discrimination index must satisfy some desiderable properties which are analogous to those defined in poverty analysis.

Del Río et al. (2011) agree with the distributional approach by Jenkins and they employ the family of indices by Foster, Greer and Thorbecke (1984) (FGT) (originally proposed for poverty analysis) for the study of wage discrimination:

$$D_\alpha = \frac{1}{n_\mathrm{F}} \sum\nolimits_{i \in P} \left(\frac{\widehat{R}_{\mathrm{F}i} - \widehat{W}_{\mathrm{F}i}}{\widehat{R}_{\mathrm{F}i}} \right)^\alpha, \quad \alpha \geq 0 \tag{2}$$

where n_F is the number of the women in the sample, $\widehat{R}_{\mathrm{F}i}$ is the expected wage which a woman would receive if she were not discriminated, $\widehat{W}_{\mathrm{F}i}$ is the unadjusted expected wage, P is the set of labels identifying discriminated women, i.e. women for whom $\widehat{R}_{\mathrm{F}i} - \widehat{W}_{\mathrm{F}i} > 0$. This index summarizes the distribution of individual discrimination, defined as $\widehat{R}_{\mathrm{F}i} - \widehat{W}_{\mathrm{F}i}$, in a single measure. The parameter α can be interpreted as an aversion parameter to discrimination: the larger is its value, the harsher is the penalty which the index attaches to a transfer of discrimination from a undiscriminated woman to a discriminated one. When $\alpha = 0$, the index is a head-count ratio of discriminated women, namely the share of discriminated women; when $\alpha > 0$ the index measures the intensity of discrimination.

The gender wage differential is determined by gender differences in productivity (which are related to human capital endowments), wage discrimination and occupational segregation. Wage discrimination occurs when two equally productive workers are paid a different amount for the same job. Occupational segregation occurs when women and men are differently distributed among occupations[1]; if women are more concentrated in low-paid occupations than men, this contributes to lowering the mean female wage. Occupational segregation can be due to occupational discrimination, that is the discriminatory behavior practised by employers, or be determined by personal preferences for a particular job.

In many analyses regarding the gender pay gap, the distribution of male and female among occupations is exogenously given, in the sense that it is not held to be generated by a discrimination process, thus masking an important source of discrimination. In this paper we will propose a methodology to separately evaluate the impact of wage discrimination and that of occupational discrimination, adopting the distributional approach by Del Río et al. (2011), which hinges on the FGT class of indices. In order to disentangle the two sources of discrimination, we need to evaluate the probability distribution of every female worker to be employed among occupations if she were treated as a man. A multinomial probit model will be separately estimated by sex to provide such information.

The remainder of this paper is organized as follows: Sect. 2 will review some basic concepts regarding segregation, occupational discrimination and their measurements; Sect. 3 will present our method; Sect. 4 will outline an empirical application on the Italian labour market data; the final section contains concluding remarks.

[1]For a review of the theories relating to occupational segregation by sex see Blau and Jusenius (1976) and Anker (1997).

Segregation and Occupational Discrimination

Whilst female workers are confined to a limited set of occupations or sectors of economic activity, segregation represents a waste of human resources and an aspect of inefficiency in the labour market. It could, therefore, be said that the focus of labour research should be on equal opportunities rather then on market results only. Thus we think it is appropriate to disentangle the concept of segregation *tout-court* from that of occupational discrimination.

The difference in the distribution of men and women among occupations is measured by indices of segregation, which summarize how much the observed configuration departs from a proportional representation of the two sexes. The most common used segregation measure is the classic segregation index by Duncan and Duncan (1955) (D&D), also known as the *index of dissimilarity*, which is defined as:

$$D = (1/2) \sum_{j=1}^{k} \left| (M_J / M) - \left(F_j / F \right) \right| \tag{3}$$

where M_J and F_j are the number of men and women respectively in occupation $j = 1, 2, \ldots, k$, and N_{M} and N_{F} are the number of male and female employees respectively. The D&D index is zero when the relative distributions of the two sexes are equal. When all men or women are concentrated into a single occupation, the index takes on the value of one. The index has a convenient interpretation: its value represents the share of women or men who are obliged to change occupation to eliminate segregation.

The D&D index and other segregation indices (Moir and Selby 1979; Karmel and MacLachlan 1988; Hutchens 2004) do not provide a measure of occupational discrimination, because they do not control for workers' personal characteristics. Indeed, segregation can be due to differences in human capital endowment, making it more likely for a particular gender to be employed in, for example, high status professions rather than unskilled jobs. Instead, occupational discrimination is a phenomenon which causes gender biases in hiring and promotion (Chzhen 2006) and it cannot be explained by strictly labour market factors.

A straightforward estimation strategy for occupational discrimination hinges on the theoretical distribution of women among occupations which would prevails if each woman in the sample had the same occupational attainment probability distribution, conditional on her characteristics, of a male worker. The impact of occupational discrimination on segregation can be measured via the comparison between the actual level of segregation and the case of free-from-discrimination occupational distribution.

Occupational attainment models employed in labour econometrics are models with qualitative dependent variable (Long 1997): the multinomial logit (or probit) model (Theil 1969), the conditional logit model (McFadden et al. 1968; McFadden 1974) and the ordered probit model (for a general discussion on the latter, see Greene 2003). We have used a multinomial logit model inthe method proposed

in this paper, according to which the estimated probability $\widehat{p}_{ij}$ to be employed in occupation j of a worker i of sex $\mathrm{S} = \mathrm{M}, \mathrm{F}$ and individual characteristics vector $\boldsymbol{X}_{\mathrm{S}i}$ is

$$\widehat{p}_{\mathrm{S}ij} = \frac{\exp\left(X_{\mathrm{S}i}\,\widehat{\gamma}_{\mathrm{S}j}\right)}{1 + \sum_{h=2}^{k} \exp\left(X_{\mathrm{S}i}\,\widehat{\gamma}_{\mathrm{S}h}\right)}, \quad \mathrm{S} = \mathrm{M}, \mathrm{F} \tag{4}$$

where $\widehat{\boldsymbol{\gamma}}_{\mathrm{S}j}$ $(j = 1, 2, \ldots, k)$ are estimated parameters with $\widehat{\boldsymbol{\gamma}}_{\mathrm{S}0}$ arbitrarily set to $\mathbf{0}$.

The estimated share of women in occupation j, if the labour market treated them as they were men, is

$$\widehat{F}_j = \sum_{i=1}^{N_{\mathrm{F}}} \frac{\exp\left(X_{\mathrm{F}i}\,\widehat{\gamma}_{\mathrm{M}j}\right)}{1 + \sum_{i=1}^{N_{\mathrm{F}}} \exp\left(X_{\mathrm{F}i}\,\widehat{\gamma}_{\mathrm{M}h}\right)} \tag{5}$$

which can be used to estimate the adjusted-for-discrimination D&D index:

$$D' = (1/2) \sum_{j=1}^{k} \left| (M_J/M) - \left(\widehat{F}_j/F\right) \right| \tag{6}$$

In empirical analysis, the D' index can be commented upon as a measure of segregation, which can be explained by differences in endowments (Brown et al. 1999) or compared with the unadjusted D index in evaluating the impact of occupational discrimination (Chzhen 2006; Miller 1987). Another estimation approach to estimating the impact of occupational discrimination, one which combines the D&D index with the multinomial logit model, is provided by Kalter (2000).

The Brown–Moon–Zoloth (B–M–Z) decomposition (1980) is an appropriate procedure for evaluating the impact of occupational segregation (explained and unexplained by individual characteristics) on the gender wage differential. It basically decomposes the wage gap in four parts: the explained (*EW*) and the unexplained (*UW*) by individual characteristics of the within-occupation wage differential, and the explained (*EO*) and the not-explained (*UO*) by individual characteristics of the between-occupation wage differential. The *UW* component can be interpreted as wage discrimination, while the *UO* component as occupational discrimination. The B–M–Z decomposition is based on separate-by-sex estimates for the parameters of a multinomial logit models for occupational attainment $(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}j},\ \mathrm{S} = \mathrm{M}, \mathrm{F},\ j = 1, 2, \ldots, k)$ and on separate-by-sex-and-occupation estimates for the parameters of $k \times 2$ within-occupation wage regression models $(\widehat{\boldsymbol{\beta}}_{\mathrm{S}j},\ \mathrm{S} = \mathrm{M}, \mathrm{F},\ j = 1, 2, \ldots, k)$. The decomposition is given by:

$$\begin{aligned} \overline{W}_{\mathrm{M}} - \overline{W}_{\mathrm{F}} = {} & \underbrace{\sum_{j=1}^{k} P_{\mathrm{F}j} \left(\overline{\boldsymbol{Z}}_{\mathrm{M}j} - \overline{\boldsymbol{Z}}_{\mathrm{F}j}\right) \widehat{\boldsymbol{\beta}}_{\mathrm{M}j}}_{EW} + \underbrace{\sum_{j=1}^{k} \overline{\boldsymbol{Z}}_{\mathrm{M}j} - \widehat{\boldsymbol{\beta}}_{\mathrm{M}j} \left(P_{\mathrm{M}j} - P'_{\mathrm{F}j}\right)}_{EO} \\ & + \underbrace{\sum_{j=1}^{k} P_{\mathrm{F}j}\, \overline{\boldsymbol{Z}}_{\mathrm{F}j} \left(\widehat{\boldsymbol{\beta}}_{\mathrm{M}j} - \widehat{\boldsymbol{\beta}}_{\mathrm{F}j}\right)}_{UW} + \underbrace{\sum_{j=1}^{k} \overline{\boldsymbol{Z}}_{\mathrm{M}j}\, \widehat{\boldsymbol{\beta}}_{\mathrm{M}j} \left(P'_{\mathrm{F}j} - P_{\mathrm{F}j}\right)}_{UO} \end{aligned} \tag{7}$$

where P_{Mj} and P_{Fj} are the actual proportions of men and women respectively in occupation j, P'_{Fj} is the estimated adjusted proportion of female workers in occupation j, calculated using (5), and $\overline{\boldsymbol{Z}}_{Mj}$ and $\overline{\boldsymbol{Z}}_{Fj}$ are the vectors of male and female mean individual characteristics respectively of workers in occupation j.

Measuring Wage Discrimination and Occupational Discrimination in the Distributional Approach

According to (Cain et al. 1986), the variables held constant in the statistical model, which is used to measure discrimination, should not be determined by the process of discrimination under examination. When occupational dummies are used in the B–O decomposition, gender differences in the distribution of workers among occupation are not justified by an occupational attainment model and the analysis thus ignores occupational discrimination. Furthermore the inclusion of occupational dummies in the wage equation is a questionable issue: while their exclusion allow for accounting for occupational discrimination, this estimation strategy, however, penalizes the accuracy of the model which explains wage (Miller 1987). Solberg (2005) claims that including dummy variables for occupation is not an adequate control and many authors found that the inclusion of occupational dummies in wage regressions reduces the unexplained component (Blau and Ferber 1987; Kidd and Shannon 1996). The B–M–Z decomposition addresses these methodological issues, but it does not take into account any distributional aspect of discrimination.

Our approach attempts to combine various features of the B–M–Z decomposition and the distributional approach by Del Río et al. (2011) in providing two separate measures for wage discrimination and occupational discrimination, which are distribution-sensitive.

Following (Brown et al. 1980), we first estimate two logit multinomial occupational attainment model with k occupations, separately by sex, using $\boldsymbol{X}_{Mi}$ and $\boldsymbol{X}_{Fi}$ individual characteristics vectors for men and women respectively. The two estimated multinomial model provide us with k estimated vectors of parameters $\widehat{\boldsymbol{\gamma}}_{Mj}$ for men and k estimated vectors $\widehat{\boldsymbol{\gamma}}_{Fj}$ for women. Thereafter, we use these estimates to assess the probability of a woman with characteristics $\boldsymbol{X}_{Fi}$ to be employed in occupation j if she were evaluated by the labor market as a man:

$p'_{Fij} = \frac{\exp\left(X_{Fi}\widehat{\gamma}_{Mj}\right)}{1+\sum_{h=2}^{k}\exp\left(X_{Fi}\widehat{\gamma}_{Mh}\right)}$, $S = M, F$. We also estimate the following lognormal wage equations, separately by sex and occupation, using individual characteristics $\boldsymbol{Z}_{Mi}$ for men and $\boldsymbol{Z}_{Fi}$ for women:

$$\log W_{Si} = Z_{Si}\beta_S + \varepsilon_{Si}, \varepsilon_{Si} \sim N\left(0; \widehat{\sigma}_S^2\right), \quad S = M, F$$

resulting in k OLS estimated vectors $\widehat{\boldsymbol{\beta}}_{Mj}$ and k analogous vectors $\widehat{\boldsymbol{\beta}}_{Fj}$. We estimate the female expected wage, which is adjusted for discrimination and conditioned to

being employed in occupation j, as $\exp\left(Z_{\mathrm{F}i}\widehat{\beta}_{\mathrm{M}j}+\widehat{\sigma}^2_{\mathrm{M}}\right)$.[2] The estimated parameters are used to predict the expected wage in absence of occupational discrimination for each woman:

$$\widehat{U}_{\mathrm{F}i}=\sum\nolimits_{j=1}^{k}\left[\frac{\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{M}j}\right)}{1+\sum_{h=2}^{k}\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{M}h}\right)}\exp\left(Z_{\mathrm{F}i}\widehat{\beta}_{\mathrm{F}j}+\frac{\widehat{\sigma}^2_{\mathrm{F}}}{2}\right)\right] \tag{8}$$

which is obtained by using the estimated male parameters in the occupational attainment model and the estimated female parameters in each within-occupation wage model.

By using the estimated female parameters in the occupational attainment model and the estimated male parameters in each within-occupation wage model, we obtain the expected wage for each woman in the absence of wage discrimination:

$$\widehat{R}_{\mathrm{F}i}=\sum\nolimits_{j=1}^{k}\left[\frac{\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{F}j}\right)}{1+\sum_{h=2}^{k}\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{F}h}\right)}\exp\left(Z_{\mathrm{F}i}\widehat{\beta}_{\mathrm{M}j}+\frac{\widehat{\sigma}^2_{\mathrm{M}}}{2}\right)\right] \tag{9}$$

Finally, we calculate the unadjusted expected wage as

$$\widehat{W}_{\mathrm{F}i}=\sum\nolimits_{j=1}^{k}\left[\frac{\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{F}j}\right)}{1+\sum_{h=2}^{k}\exp\left(X_{\mathrm{F}i}\widehat{\gamma}_{\mathrm{F}h}\right)}\exp\left(Z_{\mathrm{F}i}\widehat{\beta}_{\mathrm{F}j}+\frac{\widehat{\sigma}^2_{\mathrm{F}}}{2}\right)\right] \tag{10}$$

The distributional index of occupational discrimination we have proposed is obtained by using the FGT class of indices, where the role of "poverty line" is assumed by $\widehat{U}_{\mathrm{F}i}$:

$$\widehat{D}^{\alpha}_{\mathrm{O}}=\frac{1}{n_{\mathrm{F}}}\sum\nolimits_{i\in P_{\mathrm{O}}}\left(\frac{\widehat{U}_{\mathrm{F}i}-\widehat{W}_{\mathrm{F}i}}{\widehat{U}_{\mathrm{F}i}}\right)^{\alpha},\quad \alpha\geq 0 \tag{11}$$

where the set P_{O} identifies the women for whom $\widehat{U}_{\mathrm{F}i}-\widehat{W}_{\mathrm{F}i}>0$ (that is, the women which can be considered *discriminated* in the occupational sense) and α can be interpreted as an aversion parameter to occupational discrimination.

[2] Remember that if $\log W_{Si}\sim N\left(\boldsymbol{Z}_{Si}\boldsymbol{\beta}_S;\widehat{\sigma}^2_S\right)$ then $W_{Si}\sim \log N\left(\boldsymbol{Z}_{Si}\boldsymbol{\beta}_S;\widehat{\sigma}^2_S\right)$, thus $E\left(W_{Si}\right)=\exp\left(\boldsymbol{Z}_{Si}\boldsymbol{\beta}_S+\widehat{\sigma}^2_S\right)$. The estimator $\exp\left(\boldsymbol{Z}_{Si}\widehat{\boldsymbol{\beta}}_S+\widehat{\sigma}^2_S\right)$ is biased but consistent for $E(W_{Si})$.

Table 1 Indices of occupational discrimination $\widehat{D}_{\mathrm{O}}^{\alpha}$ and wage discrimination $\widehat{D}_{\mathrm{W}}^{\alpha}$ for different values of aversion parameter α calculated for Italy and Italian regions

	$\widehat{D}_{\mathrm{O}}^{\alpha}$				$\widehat{D}_{\mathrm{W}}^{\alpha}$			
α	North	Center	South	Italy	North	Center	South	Italy
0	0.132	0.037	0.004	0.082	0.993	0.987	0.973	0.987
1	0.004	0.000	0.000	0.002	0.154	0.140	0.124	0.144
2	0.000	0.000	0.000	0.000	0.028	0.024	0.020	0.025
3	0.000	0.000	0.000	0.000	0.005	0.005	0.004	0.005

Source: Authors' calculations using the Italian Eu-Silc 2006 data

Our distributional index of wage discrimination is given by:

$$\widehat{D}_{\mathrm{W}}^{\alpha} = \frac{1}{n_{\mathrm{F}}} \sum\nolimits_{i \in P_{\mathrm{W}}} \left(\frac{\widehat{R}_{\mathrm{F}i} - \widehat{W}_{\mathrm{F}i}}{R_{\mathrm{F}i}} \right)^{\alpha}, \quad \alpha \geq 0 \tag{12}$$

where the set P_{W} identifies the women for whom $\widehat{R}_{\mathrm{F}i} - \widehat{W}_{\mathrm{F}i} > 0$ (that is, the women who can be considered purely-wage-discriminated) and α can be interpreted as an aversion parameter to wage discrimination.

Empirical Analysis

We employed our distributional indices to analyze gender discrimination in Italy, using the Eu-Silc Italian data for 2006. The sample under consideration comprised employees (minimum age 16-years old), who were in receipt of a paid work when interviewed. The sample included 8,333 men and 6,677 women. Eight of the nine occupations of the Isco-88 (COM) one-digit classification were considered in our analysis, excluding the armed forces (the exclusion is due to the low number of women in this category). Variables used for the multinomial logit models were: number of years in education, years of work experience and dummy variables for the region of residence (the north, center or south of Italy). Variables used for the lognormal wage equations varied from occupation to occupation, being selected according to tests of significant for regression coefficients; they were generally the same as those used in the multinomial models plus worked hours in a week and economic activity. In calculating our discrimination indices, we use different values of the parameter α to provide discrimination evaluations at different levels of aversion to discrimination (the interpretation is straightforward only when $\alpha = 0, 1$). The results are shown in Table 1 below.

These results demonstrate that 98.7 % of Italian women suffer wage discrimination, while women suffering occupational discrimination are only 8.2 %. A higher value for the parameter α, the more the index reflects aversion to discrimination. Discrimination is more marked in the north of Italy but differences between the various regions do not seem to be significant for higher values of α. The ranking

of the evaluation of occupational and wage discrimination for Italian regions does not change for different values of α, thus providing a clear picture of the two discrimination forms. We further demonstrated that wage discrimination in Italy is more significant than occupational discrimination, thus providing us with an interesting interpretation of the gender pay gap.

Conclusions

The classic approach to measure discrimination, given by decomposition techniques at mean values of individual characteristics, can be considered as an approximate way to summarize individual discrimination. Indeed, it does not take into account various important properties which would characterize an effective discrimination index, such as, for example, the transfer principle. Instead, the distributional approach focuses its attention on the entire distribution of discrimination and satisfies desiderable properties which are analogous to those commonly used in poverty analysis.

Another issue in analyzing labour discrimination is the controlling for individual characteristics which determine the probability to be employed in an occupational category. The (Brown et al. 1980) decomposition gives a well-founded estimation strategy for this type of control, but relies on an evaluation at mean values of individual characteristics.

Our approach is based on an occupational attainment model, similar to that of (Brown et al. 1980), and on estimates for the expected wage, as adjusted for occupational discrimination and the expected wage, as adjusted for wage discrimination. We measured two forms of individual discrimination (of occupational and purely-wage type) and aggregate the corresponding distributions using the Foster et al. (1984) class of indices; the latter were originally used in poverty analysis and also employed in discrimination analysis by Del Río et al. (2011). Thus, we could provide two separate measures of wage discrimination and occupational discrimination.

The empirical analysis which we performed for the Italian labour market demonstrated that wage and occupational discrimination are quite different in their extent and intensity. This fact can yield important information regarding the functioning of the Italian labour market, guiding policy makers towards specific areas of intervention in gender issues.

We will conclude by outlining several theoretical challenges. Every discrimination analysis relies on the occupational detail chosen. We use the Isco-88 Com classification of occupations at a very aggregated detail, and we are aware that results could change in accordance with a different occupational detail. Furthermore, international standards classification of occupations can lead to a segregation evaluation which depends on the logic of the classification itself and, therefore, another classifications could be useful in future research.

A final consideration must be mentioned, regarding the meaning of segregation which cannot be explained by individual characteristics. As segregation can be due

to employees' individual preferences (in addition to occupational discrimination being practiced by employers), it may not be clear from ordinary empirical analysis how much of the not-explained segregation can be due to discrimination. Little research currently exists regarding the estimation strategy in providing separate measures of the impact of the two phenomena and this could lead the way to future research.

References

Anker, R.: Theories of occupational segregation by sex: an overview. Int. Labour Rev. **136**(3), 315–339 (1997)

Blau, F., Ferber, M.A.: Discrimination: empirical evidence from the United States. Am. Econ. Rev. Pap. Proc. **77**(2), 316–320 (1987)

Blau, F.D., Jusenius, C.L.: Economists' approaches to sex segregation in the labor market: an appraisal. Signs **1**(3), 181–199 (1976)

Blinder, A.: Wage discrimination: reduced form and structural estimates. J. Hum. Resour. **8**(4), 436–455 (1973)

Brown, R.S., Moon, M., Zoloth, B.S.: Incorporating occupational attainment in studies of male–female earnings differentials. J. Hum. Resour. **15**(1), 3–28 (1980)

Brown, J.B., Pagán, J.A., Rodríguez-Oreggia, E.: Occupational attainment and gender earnings differentials in Mexico. Ind. Labor Relat. Rev. **53**(1), 123–135 (1999)

Cain, G.G.: The economic analysis of labor market discrimination: a survey. In: Ashenfelter, O., Laynard, R. (eds.) Handbook of Labor Economics, vol. 1. Elsevier Science Publisher BV, Amsterdam (1986)

Chzhen, Y.: Occupational gender segregation and discrimination in Western Europe. Paper prepared for EPUNet conference 2006, Barcellona, 8–9 May 2006

Cotton, J.: On the decomposition of wage differentials. Rev. Econ. Stat. **70**(2), 236–243 (1988)

Del Río, C., Gradín, C., Cantó, O.: The measurement of gender wage discrimination: the distributional approach revisited. J. Econ. Inequal. **9**(1), 57–86 (2011)

Duncan O., Duncan B.: A Methodological Analysis of Segregation Indexes. Am. Sociol. Rev. **20**(2), 210–217 (1955)

Foster, J., Greer, J., Thorbecke, E.: A class of decomposable poverty measures. Econometrica **52**(3), 761–766 (1984)

Greene, W.H.: Econometric Analysis. Prentice Hall, Upper Saddle River (2003)

Hutchens, R.: One measure of segregation. Int. Econ. Rev. **45**(2), 555–578 (2004)

Jenkins, S.P.: Earnings discrimination measurement: a distributional approach. J. Econ. **61**(1), 81–102 (1994)

Kalter, F.: Measuring Segregation and Controlling for Independent Variables. Working paper n. 19 of the Mannheimer Zentrum für Europaische Sozialforschung

Karmel, T., MacLachlan, M.: Occupational sex segregation. Increasing or decreasing? Econ. Rec. **64**(3), 187–195 (1998)

Kidd, M.P., Shannon, M.: Does the level of occupational aggregation affect estimates of the gender wage gap? Ind. Labor Relat. Rev. **49**(2), 317–329 (1996)

Long, J.S.: Regression Models for Categorical and Limited Dependent Variables. Sage, Thousand Oaks (1997)

McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.) Frontiers in Econometrics, pp. 105–142. Academic, New York (1974)

McFadden, D.: The revealed preferences of a government bureaucracy. Economic Growth Project, Technical Report no. 17. University of California (1968)

Miller, P.W.: The wage effect of the occupational segregation of women in Britain. Econ. J. **97**(388), 885–896 (1987)

Moir, H., Selby, S.J.: Industrial segregation in the Australian labour market. J. Ind. Relations **21**(3), 281–291 (1979)
Neumark, D.: Employers' discriminatory behavior and the estimation of wage discrimination. J. Hum. Resour. **23**(3), 279–295 (1988)
Oaxaca, R.: Male–female wage differential in urban labor markets. Int. Econ. Rev. **14**(3), 693–709 (1973)
Oaxaca, R.L., Ransom, M.R.: On discrimination and the decomposition of wage differentials. J. Econ. **61**(1), 5–21 (1994)
Reimers, C.W.: Labor market discrimination against Hispanic and black men. Rev. Econ. Stat. **65**(4), 570–579 (1983)
Solberg, E.J.: The gender pay gap by occupation: a test of the crowding hypothesis. Contemp. Econ. Policy **23**(1), 129–148 (2005)
Theil, H.: A multinomial extension of the linear logit model. Int. Econ. Rev. **10**(3), 251–259 (1969)

Statistics and Economics: A Complex Relationship

Alessandro Roncaglia

Abstract

The relationship between economics and statistics is discussed within the perspective of the history of economic thought. A common origin is traced to William Petty's political arithmetic. His inductive method, as well as its roots in Bacon, is illustrated; it implies that importance be attributed to a first step of analysis consisting in organising a conceptual structure—a vision of the world. Petty's idea, derived from Galileo, of natural laws embedded in reality, is then discussed recalling Adam Smith's criticisms of it and his own method of the rhetoric of scientific debate. Similar conclusions are derived from Keynes's approach to the theory of probability, with his critiques to the frequentist and subjectivist approaches, his notion of the weight of the argument and his theory of groups. The concluding remarks refer to the connection between the teaching of statistics and economics in the Faculty of Statistics.

The connections between economics and statistics are close but complex; for a better understanding they need to be considered within the perspective of the history of thought.

This does not mean going back in time to remote origins. Attempts to trace back the origins of our disciplines may lead us as far back as we like, but the exercise in retrospective exploration rarely proves interesting. What is worth recalling here is the fact that for a long time and in an important line of research—that of the political arithmeticians—no distinction was made between the figure of the economist and that of the statistician (nor even of the demographer), for they all converged in the

A. Roncaglia (✉)
Società Italiana degli Economisti and Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy
e-mail: alessandro.roncaglia@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_23,

same figures: William Petty and John Graunt around the mid-seventeenth century, and then Gregory King and Charles Davenant and so on, to Patrick Colquhoun with his *Treatise on the wealth, power and resources of the British Empire* of 1814, taking in on the way the famous missing appendix of the *Essay on the nature of trade in general* by Richard Cantillon (1755). Political arithmetic was the term used by Petty himself and later by his successors to denote their researches based on quantitative estimation of the main characteristics of the contemporary societies.

Attempts to identify a separate origin for one of our disciplines, attributing to John Graunt alone the origin of demography, come up against the fact that William Petty clearly exerted a strong influence on his friend's work, to the extent that it has even been conjectured that the *Natural and political observations upon the bills of mortality of the City of London*, published in 1662, were in fact written by Petty, anxious to have his friend admitted to the Society for the Advancement of Learning (or Royal Society, as it is known today).[1] This is most probably a far-fetched hypothesis, but Petty did have an influence, and a great one; in any case, it is certainly a mistake to isolate Graunt's work from the broader context of political arithmetic. This is, of course, just one episode, but it helps to remind us that political arithmetic represented one broad line of research.

There can be no denying that subsequently the various disciplines followed largely independent courses for a long period of time. It was time when, more generally speaking, in the natural sciences as in the social sciences, segmentation of research went hand-in-hand with scientific progress, and indeed has often been identified with it. In the economics field we have industrial economists, economists of labour, money, energy and tourism, public finance, historians of economic thought and so on. Similar divisions also exist in the field of statistics, albeit with possibly less clear-cut demarcations. The SIS (Società Italiana di Statistica) counts among its members probabilists and demographers, market researchers and operations researchers, and so forth.

We should distinguish between segmentation in the fields of research and specialisation among researchers. The latter is inevitable despite the fact that often in the course of a lifetime's research there is time to address various issues and acquire a range of specialisations. Segmentation, on the other hand, has to do above all with the professionalisation of research, and thus with the organisation of academic careers, driving us, for example, to take great pains over redefining chair groupings. Segmentation can be very harmful, as for example when it results in relegating fields of research on the borderline between sectors to a limbo, as is the case today with the history of economic thought. But even specialisation itself, albeit undoubtedly necessary within certain limits, can only work well as long as we do not lose sight of the connections with the foundations of fields of research that may seem to be somewhat remote from the field chosen to work in. To take just a couple of examples, the labour economist cannot afford to be totally unaware of political

[1]This thesis was argued in particular by Charles Hull, in his introduction to his edition of Petty's writings, Petty (1899), where he also reprinted the *Natural and political observations*.

organisations or labour law, while the demographer must have some acquaintance with Markov chains and economics.

Taken to extremes, both specialisation and segmentation are harmful. Twenty years ago Giacomo Becattini, then president of the Società Italiana degli Economisti, had already warned of "specialists of the left big toe" who lose sight of the connections with the rest of the foot, not to mention the rest of the human body. Since then segmentation and, above all, specialisation have continued to forge ahead.

For example, it can happen—in fact it does happen quite often—for an industrial economist unacquainted with the history of thought to repeat unawares the very mistakes made by Marshall that Sraffa had already pointed out in 1925. Indeed, in some cases—and in particular with reference to the separation between macro and microeconomics—the division between fields of research may play a crucial role in defending against criticisms which could otherwise prove unanswerable, for it allows macroeconomists to ignore the findings of the debate on capital theory, traditionally included in the field of value theory, and thus of microeconomics.

If relations between the different fields of research within economics (or, I imagine, within statistics) are already so difficult, the difficulties are obviously all the greater if we go on to consider relations between statisticians and economists.[2] The distinction between the two fields—statistics and economics—is clear enough today, and there is no need to go into precise definitions (which could, indeed, complicate matters). Although I graduated in statistics, I cannot claim to be considered a statistician, nor even "also" a statistician, as well as being an economist. But if we look back to the period before the professionalisation of our research fields, we see a rather different situation. I had heard of William Petty and the tradition of political arithmeticians from Vittorio Castellano, who was my professor of statistics, before I was told about them by Paolo Sylos Labini and Piero Sraffa, my masters in economics. Each of them referred to different aspects of a personality as multifaceted as that of Petty, who had also been among other things a cartographer, a physician, a professor of music and a nautical engineer. What is, however, certain, is that the statistician and the economist cohabited with no distinction in Petty's political arithmetic.[3]

To bring Petty's position into clearer focus, we will take a step back. As Petty himself explicitly recognised, behind his method—the method of political arithmetic—lay Bacon, whose teaching in this respect can be summed up in a celebrated passage:

> the men of experiment are like the ant; they only collect and use: the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes a middle course; it gathers its material from the flowers of the garden and of the field, but transforms and digests it by a power of its own. Not unlike this is the true business of the philosophy; for it neither relies solely or chiefly on the power of the mind, nor does it take the matter which

[2] Cf. Roncaglia (2011).

[3] Cf. Roncaglia (1977).

> it gathers from natural history and mechanical experiments and lay it up in the memory whole, as it finds it; but lays it up in the understanding altered and digested.[4]

Here Bacon contrasts the inductive method—a blend of empiricism and rationalism—with the syllogistic-deductive method of the Aristotelian tradition and the Renaissance tradition of pure empiricists (technicians and alchemists). The reasoning—the theoretical elaboration—neither precedes nor follows the collection and processing of data, but accompanies it. Not only are the data organised within a sort of "preanalytic vision", as Schumpeter called it, i.e. within a broad outline conception of the object of study, but it is also necessary to piece together a network of interrelated concepts forming the framework within which the collection of data takes place. The latter operation may prove complex, as the experts examining national accounts know all too well, but it is of prime importance; Schumpeter himself saw it as the first step, after the "preanalytic vision", in the work of research in the field of the social sciences.[5]

To take an example, unemployment data are collected on the basis of precise definitions that may be modified in the course of time and generally differ from country to country; nevertheless, behind these definitions lies a theoretical approach—the neoclassical or marginalist one—according to which unemployment measures the distance from full employment equilibrium, defined in terms of equality between labour demand and supply. On the other hand, in classical and Sraffian theory there is no such thing as the notion of a full employment equilibrium and, correspondingly, there is no notion of unemployment or rate of unemployment, nor are there any attempts to measure this magnitude. In classical theory, rather, the central concept lies in the levels or rates of activity or employment.[6]

Crude empiricism in which all conceptual construction is eschewed can go no further than collecting disordered information on realities. However, it is in fact necessary to impose some order on the collection of information unless we are prepared to make do with a one-to-one scale description of reality—a feat that is not only practically impossible but also useless, for it would offer no help in getting the bearings of a situation, which is precisely what the social scientist sets out to do.

Alongside his condemnation of crude empiricism, Bacon comes up with at least equally radical indictment of the syllogists—scholars set on confining research activity to mere deductive logic. His reference here, as we have seen, is to the tradition of Scholastic philosophy, but his critique might equally be taken to apply to attempts to reduce the various fields of social sciences to applications of mathematics—attempts that have in fact been stepped up over the last few years, despite the warnings of Fuà, Becattini, Lombardini, Sylos Labini and others in a celebrated letter published in the daily newspaper *Repubblica* on 30 October 1988.

[4] Bacon (1620), Book 1 of the *Aphorisms*, No. 95.

[5] Cf. Schumpeter (1954), pp. 41–42.

[6] Cf. Roncaglia (2006).

With his political arithmetic Petty was working within the methodological framework indicated by Bacon, but there was an important additional element, namely the influence of Hobbes's materialist sensism. When he states "Pondere mensura et numero Deus omnia fecit" (citing a famous passage from the Bible, the Book of Wisdom, xi. 20), Petty makes it clear that he sees in political arithmetic an adequate tool not only to describe reality but also to represent it theoretically, precisely because, according to the materialistic–mechanical tradition pursued by Galileo and Hobbes, reality itself has quantitative structure. Let us recall in this connection the famous passage in Galileo:

> this great book which is open in front of our eyes – I mean the Universe – [. . .] is written in mathematical characters.[7]

In other words, for Petty it is not only a matter of *surveying* and *describing* reality in terms of number, weight and measure, i.e. in quantitative terms, but also of expressing oneself in those terms in the attempt to *interpret* reality, identifying its salient characteristics, precisely because the inner structure of reality "is written in mathematical characters", for the physical sciences as much as for the sciences of the human body or the social sciences. The task of the scientist is therefore to *discover* these laws, in the etymological sense of removing the cover that hides them, identifying them beneath the covering of those multifarious contingencies that complicate the world—without, however, changing its intrinsic nature.

This eminently clear-cut methodological conception attended upon the birth of the modern sciences and their early developments, from the physics of Newton to the chemistry of Laplace, from anatomy to politic arithmetic. Contrasting with it, about a century later, we have Adam Smith's approach, stated with a degree of emphasis in his celebrated *Wealth of Nations*: "I have no great faith in political arithmetick".[8] Often this position has been interpreted as mistrust in the rough and ready methods applied in estimating statistical data by Petty and his followers, but it was not simply this (or, at least, not only this). Undeniably, Petty's data were assembled in a decidedly rudimentary way, ingenious as it often proved. Since then considerable progress has been made in the methods of collecting statistical data. Nevertheless, even today, albeit with all due caution, many make use of the estimations of the last of the political arithmeticians, the late Angus Maddison, on population and income trends from the year 0 of the modern age to today,[9] and many worthy economic historians and statisticians are engaged in the labour of reconstructing the time series with no hope of achieving the degree of precision we may take to be, averagely speaking, guaranteed for the data produced by Istat (the Italian Statistical Institute) today. However, the point Smith was making with his declaration of no confidence in political arithmetic is rather more important. It represented a methodological conception differing from Petty in at least some

[7] Galilei (1623), p. 121.

[8] Cf. Smith (1776), p. 534.

[9] Cf. Maddison (2007).

important respects, implying a critique of the idea that political arithmetic—or, in our context, statistical analysis—opens the way to the discovery of actual "laws" inherent in nature and society (and thus something different from simple statistical regularities or "stylised facts", as Nicholas Kaldor used to call them).

Smith referred to this methodological conception in the opening pages of his *History of Astronomy*, a text highly praised by Schumpeter, who challengingly considered it the only work by Smith endowed with real originality.[10] Before recounting the history of the transition from the Ptolemaic to the Copernican conception, Smith[11] explained that nature appears to us as a series of "events which appear solitary and incoherent with all that go before them, which therefore disturb the easy movement of the imagination"; in order to surmount this vexing situation men have recourse to philosophy, or scientific reflection. More precisely, philosophy is defined as "the science of the connecting principles of nature" (where again we find the conception of the sciences of nature, man and society as all one). The task of philosophy is "to introduce order into this chaos of jarring and discordant appearances", "by representing the invisible chains which bind together all these disjointed objects". Performance of this task entails construction of "philosophical systems" (just like the two different cosmological conceptions, Ptolemaic and Copernican, illustrated by Smith in the following pages of his text). These philosophical systems, Smith stresses, are "mere inventions of the imagination, to connect together the otherwise disjointed and discordant phaenomena of nature". In other words, the researcher contemplating some aspect of the world and seeking to interpret its functioning (the philosopher, in Smith's terminology) has an active role, of creation and not *discovery* of theories. Thus Smith rejects the idea of a mathematical structure for reality, as Galileo had argued for physics and astronomy, as Hobbes and subsequently Condillac with sensism had extended to the human body and as Petty and the political arithmeticians had extended to the "body politic", or in other words society.

"Philosophical systems" like the Ptolemaic or Copernican systems may be "inventions of the imagination", but they can help us to get our bearings amid the chaos of real events. However, there is clearly no possibility to *verify* the theories by demonstrating that they correspond to the intrinsic laws of nature, for this would require such laws to have an existence of their own, independently of the theories: to be inscribed in the real world, and not a creation of our thought. In comparing different philosophical systems and choosing between them there are, therefore, no straightforward, unambiguous answers.

Thus Smith anticipated the positions recently advanced by Feyerabend or McCloskey, reference being to "honest conversation", a "rhetoric" of scientific debate. More precisely, in the *Lectures on rhetoric* Smith proposed a model

[10]Cf. Schumpeter (1954), p. 182. Actually, Smith proposes theses very much like those illustrated by Thomas Kuhn (a junior colleague of Schumpeter at Harvard) 50 years ago in his successful book on *The Structure of Scientific Revolutions* (1962)

[11]Smith (1795); the passages quoted below occur on pp. 45, 46, 105.

resembling procedure in a trial at law, presenting the evidence for or against a certain thesis in order to choose which propositions to accept or reject.[12] Such ideas have a long tradition behind them, going back to the Greeks Sophists who disputed the theses advanced by Socrates and Plato on the existence of a Truth inscribed in reality which philosophical investigation should seek to unveil. By contrast, the Sophists advocated open discussion on the pros and cons of each particular thesis.

In scientific debate the rhetoric method presupposes honest behaviour on the part of all the participants in the debate. This connection between theory of knowledge and ethics is made possible by recourse to the notion of the impartial spectator, which Smith (1976) set out in his *Theory of moral sentiments*, published in 1759. Individuals evaluate their actions (in our case researchers evaluate their theories) taking the point of view of an "impartial spectator" who, aware of all the details known to them, abandons all personal interests and judges on the basis of the moral criterion developed shortly after by Kant in the *Critique of judgement*.

Apart from method, a further aspect distinguishing Smith from Petty, an undeclared but recognisable follower of Machiavelli, is precisely the ineradicable presence of an ethical element in analysis of reality, both in his method of work, as we have seen, and in his choice of problems and, indeed, in the simplifications of reality that have to be made to be able to address them at the theoretical level.

Working with statistics also remains fundamental to this methodological approach, but no longer as a means to discover and represent laws inscribed in reality, nor as a tool to verify theoretical findings. Here, rather, statistical work constitutes a fundamental tool to assemble evidence in support of this or that thesis, albeit with the understanding that the various theses can never be definitively judged true or false, but only more or less likely in the light of the knowledge so far acquired. We may follow Popper (1934) in holding that confutation is more important than confirmation in the work of scientific research, but always with the proviso that no such confutation can be taken to be definitive, either. As Popper put it, one black swan sufficed to confute the thesis that all swans are white; yet DNA analysis of black swans—impossible 80 years ago when this variety of swan was discovered in Australia—might in principle have subsequently demonstrated such a genetic distance from white swans as to force us to consider them two different species.

A position similar in many respects to Smith's was arrived at a century and a half later by John Maynard Keynes, with his *Treatise on probability* (1921). In the succession of various approaches to probability, Keynes's contribution followed upon the classic conceptualisation by James Bernoulli (1654–1705; his *Ars conjectandi* was published posthumously in 1713) and his immediate successors, including his nephew Daniel Bernoulli (1700–1782), and upon the frequentist approach expounded by, among others, John Venn (1834–1923), author of a celebrated text, *Logic of chance* (1866), and professor at Cambridge at the time Keynes was studying

[12]Cf. Smith (1983), p. 178.

there, but before the subjectivist line taken by De Finetti (1930), Ramsey (1931) and Savage (1954).[13]

As we know, in the classical definition probability is expressed as the number of favourable cases divided by the number of possible cases. This definition applies particularly well to "regular" games like roulette or dice, provided, of course, that we are thinking of a perfectly constructed roulette wheel or dice that are not loaded. In fact, the definition implies complete specification of the range of events divided into a finite number of elementary events (e.g. the six sides of the die, the thirty-seven numbers of the roulette wheel—bearing in mind that zero is also to be included), considered equally probable on the basis of the principle of insufficient reason or principle of indifference, according to which there is no reason to consider one elementary event more probable than another. The task pursued by the calculus of probability is to determine the probability of complex events, like two or seven emerging as the sum when two dice are thrown.

Here we come up against two limitations. To begin with, applying the findings of probability calculus to a concrete game implies a distinct assumption which is not always verified and which we cannot generally rely on being correct, namely that the concrete game in question is practically indistinguishable from the perfectly "regular" game analysed at the level of the theory. Secondly, and even more importantly, the vast majority of the problems that we are interested in studying from a probabilistic point of view have nothing like the same characteristics as the regular game.

On the other hand, the frequentist approach has to do with the inductive knowledge methodology. The idea is that the probability of an event is the limit to which the relative frequency of the event tends in successive observations (stochastically independent from each other) of some variable, for instance the stature of conscripts or the throw of a die, or repeated independent measures of the same magnitude, when the number of observations tends to infinity. As with the classical definition, the frequentist definition implies an objective view of probability. The objective nature of the probability statement lies in the fact that it is considered to depend on the intrinsic properties of the phenomenon under consideration, and thus in a sense derived from them.

Strictly speaking, as Richard von Mises, an exponent of the frequentist approach pointed out, it can be applied only to "collectives", or in other words successions of uniform events only differing in some observable characteristic which is the object of scrutiny, when the principle of randomness holds—when, that is, no regular sequence occurs, making it impossible to devise a successful strategy applicable to the order of sequence. Clearly, these are very restrictive conditions which should stand in the way of application of the frequentist approach to any social phenomena. More generally speaking, theoretically if we interpret the requisite of infinite length of the series of observations literally, no finite series of observations, however long, can constitute a collective. In this respect we may recall the sceptical view of

[13]For more detailed illustration of the following remarks, cf. Roncaglia (2009).

the inductive method taken by David Hume long ago, in the eighteenth century: experience of past events can play an important role in moulding our beliefs; however, it can be contradicted by subsequent events; consequently, we cannot infer from a sequence of past events generalisations applicable to future events.

According to the subjective (or "personalist") approach which was proposed around 1930 by Bruno de Finetti and Frank Ramsey independently of one another, and which gained ground after publication of the text by Savage (1954), the statement of probability is subjective in the sense that it is a state of mind, not a state of nature. Probability can be defined as the lowest betting odds one would accept on a given event. If the individual in question is indifferent to the event, then the "supply price" and "demand price" of the bet are equal, and correspond to the assessment of the probability of the event considered by that individual. The mathematical theory of probability is entrusted with the task of ensuring the logical consistency of each agent's book of bet offers, identifying arbitrage strategies should misalignments arise.

According to the Italian mathematician Francesco Paolo Cantelli (1875–1966), the field of probability calculus is made up of various subfields, for each of which one of the various approaches mentioned above will prove the most suitable: urn theory (conceived as a development of the classical theory of probability) for cases in which equally probable atomic events can be defined; the frequentist approach for matters of insurance; and the subjective betting approach for fields such as horse races. Naturally, as pointed out by De Finetti, the mathematical treatment is similar in the three cases; what Cantelli meant to stress was the different natures of the phenomena considered, implying different procedures to assemble the data upon which to perform probabilistic analysis.

Keynes, who acquired a grounding in mathematics in Cambridge, offered an original contribution to the topic in a book (Keynes 1921) which began as a fellowship dissertation but took several years to complete. His contribution, which he continued to defend in the following years, even in the face of the new subjective approach formulated by his friend Frank Ramsey, addresses three aspects: definition of probability as pertaining to the field of logic, the concept of the "weight of the argument" and the so-termed "theory of groups".

Keynes defines probability as "the degree of rational belief" that one may have in a proposition (a hypothesis) on the basis of the available evidence. Thus in itself probability is not an objective property of the phenomena under consideration, but a logical relationship introduced by the agent between the available evidence, on the one hand, and the proposition under consideration (primary proposition) on the other. The logical relationship (or secondary proposition) can differ from one agent to another due to differences in the knowledge each may have, but also to differences in intellective powers. At the same time, the probability statement retains a certain empirical correlative in the reference to the available evidence, which has the effect of a constraint on the rational observer.

In this connection we may recall the ethical element which Smith introduced into scientific debate: the agent cannot decide at will the probability to be attributed to an event (the secondary proposition), but must respect the evidence at his disposal,

seeking to infer the same evaluations of probability as anyone else might. For Keynes, then, it is not only the internal consistency of the subjective evaluations of probability of each person that counts, but also the objective correspondence with what we know; indeed, although it cannot be unambiguously specified, this objective element constitutes the dominant aspect, to the extent that we may classify Keynes's among the objective rather than subjective theories of probability.

In this way we can distinguish "rational" from "irrational" belief; the ethic of individual responsibility, in which Keynes follows Moore, leads our evaluations of probability to respect of the hard facts, as far as we can know them, since our evaluations are to be used as guide for our actions. From this point of view, the evaluation of probability must be as independent as possible from our subjective preferences. We may take the case of a doctor working on a diagnosis: to begin with, there is the effort to gather the relevant evidence regarding the patient, after which it is time to draw a conclusion that, fallible as it may be, is the best the doctor can offer, where "the best" does not mean the most optimistic one, but the one that best corresponds to the actual state of the patient. (Here it is also worth considering the concept of relevant evidence: there is no need to know everything about the patient and surrounding world, but as much as can possibly serve to evaluate the possibility of an illness and contagion, also taking into account the fact that the diagnosis has to be made as promptly as possible, or at least within a reasonable length of time, and this is a circumstance characterising the diagnoses of both the doctor and the economist. The relevant evidence has to do with the problem addressed, implying selection guided by the initial hypotheses which can then be redirected towards new aspects on the basis of the provisional evaluations as they are formulated.)

The second notion that Keynes introduces into his logic of probability is the "weight of the argument", defined as the width of the evidence upon which our probability statement is grounded. This is an additional dimension in evaluation of probability. It cannot always be expressed quantitatively, nor indeed precisely, given the logical impossibility of defining a priori the entire field of evidence which would serve to express a sure evaluation of probability. (Definition of the weight of the argument as the ratio between the evidence known to us and the full evidence can help us to understand the concept, but does not allow us to measure it, since the very notion of full evidence cannot in general be defined a priori.) Nevertheless, even a rough idea of the weight of the argument can serve to distinguish between essentially different situations. For example—and coming to a point of considerable relevance to the economic theory that Keynes was to develop in the following years—we can distinguish between the uncertainty regarding the decisions that entrepreneurs have to take on levels of investment in new plant and the decisions that financiers have to make in choosing the assets in which to invest their financial resources. The problems of everyday life are normally characterised by an intermediate degree of uncertainty, somewhere between the maximum associated with total ignorance and the minimum associated with established facts or matters of probabilistic risk (such as playing with dice deemed not to be loaded).

This is a point that needs to be borne in mind if we are to understand the third element of Keynes's theory, the theory of groups. The "group" is specified

in purely logical terms, as a set of propositions with two components: those propositions (independent of one another) that define the group as premises, and the propositions that can logically be derived from the premises. Within each "group" we can provide evaluations of probability, and the probability calculus ensures their internal consistency. The idea of the group recalls the "language games" of the later Wittgenstein (1953), the Wittgenstein of the *Philosophical Investigations*, but also the stress placed by Alfred Marshall, Keynes's master, on the use of "short causal chains".[14] Our analyses must address well-defined issues; they cannot cover overlarge fields without losing solidity.

In the light of this as well as of the critiques raised against the frequentist approach, the inductive method—and thus the use of statistical inference—calls for a considerable degree of caution. This is certainly not to say it should be rejected, for the tools of statistics are indeed of great help in getting the bearings of the available evidence, but there is the need to guard against the possibility of inferring laws of general applicability from the analysis of specific datasets.

In other words, the prospect that Keynes opens up for us lies along a tricky path between the Scylla of the uncritical empiricism of the frequentist tradition and the Charybdis of the solipsism of the subjective approach, which concentrates on the internal consistency of the system of individual beliefs while failing to take into account the fact that such beliefs guide action insofar as they refer to the realities we are faced with.

The foregoing points can be taken as a background for a few final, alas somewhat pessimistic, considerations of a personal nature on the relationship between statistics and economics and, more generally, between statistics and the social sciences in research and teaching.

When the time came for me to enrol at university, having decided to study economics I sought advice on the faculty I should apply to—economics, political sciences, social sciences, law or statistics? The most useful advice came from two sources. An elderly great-uncle, professor of commercial law, sagely advised me: "What counts are the professors you will study with, more than the subjects you study. Enrol where Sylos Labini teaches, whichever that faculty may be". So I went to meet Sylos: "To be an economist," he told me, "you need history, philosophy, mathematics and statistics. The latter are essential and hard to learn on your own; enrol in Statistics". Luckily the two recommendations coincided, and so I enrolled in Statistics.

Of course, this would not have been the most appropriate choice if I had wanted to study economics as a technical subject, to seek a job in some firm. But it was the best choice to study economics as a social science. Application of statistical methodology is impossible without a good knowledge of the field of the issues to be investigated; in fact, as we have seen, statistical inference cannot lead to scientifically definitive results. At the same time, a discipline like economics cannot be cultivated by deducing abstract theories from a priori, like Bacon's spiders,

[14] On the notion of short causal chains, cf. Roncaglia (2004).

but calls for constant comparison with reality—a scientifically serious comparison, mediated by the tools of quantitative analysis. This, I believe, applies to all the social sciences, which all need strong quantitative support and sound foundations of mathematical logic. It is a fact that has found increasing recognition throughout the world, with repercussions on the organisation of studies in the field of the social sciences. And I believe it lay behind the happy intuition of the founder of the Rome Faculty of Statistics, seeking to achieve a balance between mathematical and statistical tools and the substantive social disciplines in the order of studies.

Alas, this cultural plan—which saw the school of Sylos Labini turning out a score of full professors of economics, and top executives at the Treasury and the Bank of Italy, as well as at the World Bank and the OECD and elsewhere—has now been abandoned, with the closing first of the degree course in Statistics and Economics, then of the Department of Economics founded by Sylos Labini, to be followed by the Department of Social Sciences which we had fallen back upon, and finally the Faculty of Statistics. The idea of producing economists with a leaning towards finance, as a joint product of the preparation of actuaries, could have been, within certain limits, a useful addition to a plan for the preparation of quantitative sociologists and economists, but not a substitute for it. Moving in this direction, the tradition of statistical studies itself is doomed to steady decline. I can only hope that the other universities do not repeat the mistakes committed by the Sapienza University of Rome.

References

Bacon, F.: Novum Organum. Joannem Billium, London (1620)
Bernoulli, J.: Ars conjectandi. Thurnisiorum, Basel (1713)
Cantillon, R.: Essai sur la nature du commerce en général. Fletcher Gyles, London (1755)
Colquhoun, P.: Treatise on the Wealth, Power and Resources of the British Empire. J. Mawman, London (1814)
De Finetti, B.: Fondamenti logici del ragionamento probabilistico. Boll. dell'Unione Matematica Ital. **9**, 258–261 (1930)
Galilei, G.: Il Saggiatore. Giacomo Mascardi e Accademia dei Lincei, Roma (1623). Repr. In Opere, Ricciardi, Milano-Napoli (1953)
Keynes, J.M.: A Treatise on Probability. Macmillan, London (1921)
Kuhn, T.S.: The Structure of Scientific Revolutions. Princeton University Press, Princeton (1962)
Maddison, A.: Contours of the World Economy, 1-2030 AD. Oxford University Press, Oxford (2007)
Petty, W.: In: Hull, C. (ed.) Economic Writings, vol. 2, Cambridge University Press, Cambridge (1899)
Popper, K.R.: Logik der Forschnung. Springer, Wien (1934)
Ramsey, F.P.: The Foundations of Mathematics. Routledge and Kegan Paul, London (1931)
Roncaglia, A.: Petty: La nascita dell'economia politica. Etas Libri, Milano (1977)
Roncaglia, A.: Tasa de desempleo y tasas de empleo: ¿Categorías estadísticas o construcciones teóricas? Invest. Económica **65**, 45–61 (2006)
Roncaglia, A.: Keynes and probability: an assessment. Eur. J. Hist. Econ. Thought **16**, 489–510 (2009)
Roncaglia, A.: Macroeconomics in crisis and macroeconomics in recovery. PSL Q. Rev. **64**, 167–185 (2011)

Roncaglia, A.: Le catene causali brevi: le variazioni di Maynard su un tema di Alfred. In: Bellanca, N., Dardi, M., Raffaelli, T. (eds.) Economia senza gabbie. Studi in onore di Giacomo Becattini, pp. 379–397. il Mulino, Bologna (2004)
Savage, L.J.: The Foundation of Statistics. Wiley, New York (1954)
Schumpeter, J.: History of Economic Analysis. Oxford University Press, New York (1954)
Smith, A.: In: Bryce, J.C. (ed.) Lectures on Rhetoric and Belles Lettres. Oxford University Press, Oxford (1983)
Smith, A.: An Inquiry into the Nature and Causes of the Wealth of Nations. W. Strahan and T. Cadell, London (1776). Critical edition, Campbell, R.H., Skinner, A.S. (eds.) Oxford University Press, Oxford (1976)
Smith, A.: In: Wightman, W.P.D., Bryce, J.C. (eds.) Essays on Philosophical Subjects. T. Cadell and W. Davies, Critical edirion, London (1795). Oxford University Press, Oxford (1980)
Smith, A.: The Theory of Moral Sentiments. A. Millar, London (1759). Critical edition, Raphael, D.D., Macfie A.L. (eds.) Oxford University Press, Oxford (1976)
Sraffa, P.: Sulle relazioni fra costo e quantità prodotta. Ann. Economia **2**, 277–328 (1925)
Venn, J.: The Logic of Chance. Macmillan, London (1866)
Wittgenstein, L.: In: Anscombe, G.E.M., Rhees, R. (eds.) Philosophische Untersuchungen (with English transl., Philosophical Investigations). Blackwell, Oxford (1953)

Part V

New Methodological Developments in Educational Studies

Design, Implementation and Validation of a Questionnaire for University Teaching Evaluation

Luigi D'Ambra and Maurizio Carpita

Abstract

This paper summarizes the results of a research project supported by CNVSU in 2010. The aim of the project was, firstly, to review the questionnaire used in Italy for university teaching evaluation and, secondly, to propose a guideline for implementing this survey on the web.

Introduction

This paper presents the main results of the Project carried out in 2010, supported by CNVSU (Comitato Nazionale per la Valutazione del Sistema Universitario—National Committee for University System Evaluation) and developed by a large Research Group[1] with the aim to:

[1]The Research Group of the CNVSU Project was coordinated by Luigi D'Ambra and composed by Maurizio Carpita and Eugenio Brentari (University of Brescia), Sergio Scippacercola (University of Naples—Federico II), Pietro Amenta and Biagio Simonetti (University of Sannio), Rosaria Lombardo and Antonello D'Ambra (University of Naples II), Pasquale Sarnacchiaro (University Telma Sapienza). The final report of the Project in downloadable from the CNVSU web site: www.cnvsu.it/_library/downloadfile.asp?id=11775.

L. D'Ambra
University of Naples Federico II, Naples, Italy
e-mail: dambra@unina.it

M. Carpita (✉)
University of Brescia, Brescia, Italy
e-mail: carpita@eco.unibs.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_24,

1. Design, build and evaluate the design, build and validate the items and their rating scales in order to compose a questionnaire and the appropriate rating scale of the questionnaire for the evaluation of university teaching in Italy;
2. Prepare a guideline and its instructions (good practices) for implementing and completing the survey on the web.

The Research Group, after taking into account the structure of some standard and experimental questionnaires, and collecting data from some samples of university students, carried out a coordinated set of statistical analyses in order to evaluate the items adequacy, the possible factor structures, the construct reliability and validity and the process of measurement (rating scales).

The standard questionnaire used for the evaluation of university teaching has 15 items (CNVSU 2000) and in this experimentation has been proposed in two versions, with different rating scales:

– Version A, with items rated on a 4-point scale;
– Version B, with items rated on a 10-point scale.

The experimental questionnaire has nine items and two different rating scales:

– Version A, with items rated on two joint 4- and 10-point scales;
– Version B, with items rated on two disjoint 4- and 10-point scales.

From the statistical point of view, our aim has been twofold: (1) to identify the sub-dimensions and (2) to study the response scale of the "quality of university teaching" questionnaires. First, the results of the preliminary classical item analysis were used to direct the factor analysis (Minres method) applied in order to verify the adequacy of each item with respect to some satisfaction sub-dimensions without assumptions on the distributional properties of the items. Second, the Rasch analysis (Rating scale model) pointed out the most appropriate ordinal response category structures among those considered.

The data of the questionnaires for university teaching evaluation was collected by means of two samples of 1,000 and 500 students drawn at the Faculty of Economics of the University of Brescia and at the University of Sannio, respectively.

The Item Analysis Results

A first and classical item analysis was carried out on the data collected at the Universities of Brescia (UniBS) and Sannio (UniSN) with the four versions of the standard and experimental questionnaires for the evaluation of university teaching (CNVSU 2010, Tables 1.8 and 1.9).[2] This preliminary data analysis was conducted to direct the successive factor analysis presented in the next paragraph.

[2]At UniBS, data were collected for all the 4 questionnaires with paper and pencil; students have answered to the questions of the standard questionnaire using both the 4- and 10-point scales. At UniSN, data were collected with a web survey for two versions of the questionnaires (standard version A with 4-point scale, experimental with 10-point scale) (for non-attending students too).

For each item, the average, standard deviation, skewness and kurtosis indices of the 10-point and the 4-point scale were computed. This preliminary analysis pointed out that the distribution of all the items was significantly different from the normal. This aspect was confirmed also by the Shapiro and Wilks tests of normality. The values of skewness and kurtosis indices of all the items highlight a negative asymmetry. These distributional findings provide evidence for the non-normality of the items, and suggest that normal-theory estimation procedures may not be appropriate for examining the underlying factor structure of the items. Then, the Cronbach's α was calculated for each section of the questionnaires. The internal consistency for each section is satisfactory, except for the standard questionnaire of UniSN, whose section called "Infrastructure" exhibited a very low value ($\alpha = 0.38$). Finally, the multiple correlation indices between each item and all the others (ISC) showed that all the items are acceptable (as suggested by Nunnally and Bernstein (1994), we considered acceptable values greater than 0.30).

The analysis of linear correlations between the items of the standard questionnaire rated on a 10-point scale for the UniBS data (Table 1) leads to a first consideration about the redundancy of some items[3]: D01 and D02 (positive correlation equal to 0.70), D07 and D08 (0.80), D12 and D13 (0.68). This redundancy could allow us to eliminate some items of the questionnaire. Nevertheless, we have to emphasize that similar items placed in sequence in the questionnaire can lead to a more uniform assessments. In addition, some questions, although highly correlated to other items, may have a specific interest, so it could be desirable not to remove them. The items with the highest correlation with the overall satisfaction (D15) are D07 (0.70), D08 (0.67) and D14 (0.6).

Finally, we suggest to remove items D01 and D02, and include them again in a final questionnaire administered at the end of the academic year.

In the proposed two experimental questionnaires three items (version A) and two items (version B) seem potentially rather redundant (CNVSU 2010, Table 1.14). The linear correlation analysis between the items of the standard and the experimental questionnaire in version B (having the same rating scale), confirmed the expected associations between the items, with some marginal exceptions (for further details see CNVSU 2010, Table 1.15). Similar results were obtained for UniSN.

The Factor Analysis Results

In order to determine the model that best described the 4- and 10-point data obtained from the standard questionnaire, the minimum residual method (Minres; Harman 1960) was applied, firstly on the linear correlation matrix and, secondly, on the polychoric correlation matrix obtained using Prelis (Version 2.54). This

[3]In order to facilitate reading, Table 1 reports correlations multiplied by 100. In addition, we considered the structure of the items, by means of the polychoric correlation indices (not reported here), obtaining similar results.

Table 1 Correlation coefficients for the 10-point scale of the standard questionnaire (UniBS data)

Item	Question	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14	D15
D01	Charged study acceptable	100														
D02	Overall organization acceptable	*70*	100													
D03	Clarity of exam modality	26	32	100												
D04	Teaching timetable respected	9	13	40	100											
D05	Teacher available for explanation	15	24	46	47	100										
D06	Sufficient prior knowledge	28	23	17	5	16	100									
D07	Teacher stimulate the interest	30	34	40	17	31	35	100								
D08	Teacher exposes clearly	27	33	42	19	31	27	*80*	100							
D09	Charged study proportional to the CFU	36	31	32	26	32	32	45	45	100						
D10	Appropriate teaching materials	21	29	37	26	38	23	47	46	43	100					
D11	Supplementary activities are useful	12	23	31	17	24	15	33	34	22	27	100				
D12	Appropriate classroom for lessons	20	27	24	22	22	10	13	15	9	17	19	100			
D13	Appropriate classroom for exercitations	16	24	26	21	24	12	13	12	7	18	21	*68*	100		
D14	Interest for the topics	28	26	27	19	31	36	51	38	37	31	20	17	16	100	
D15	Overall satisfaction	33	38	45	27	39	32	*70*	*67*	47	50	34	25	21	*60*	100

second approach is justified by the inappropriateness of the methods of factor extraction (e.g., maximum likelihood or principal components) based on normal-theory estimation with ordinal data (especially for the 4-point scale) as evidenced by the previous item analysis. The Minres procedure is equivalent to the unweighted least squares common-factor analysis and was used because it does not require distributional assumptions, is very robust, and can be used with small samples and when the polychoric correlation matrix is not positive definite (Jöreskog 2003). Factor pattern matrices were examined for simple structure and interpretability.

In order to select the appropriate number of factors (the sub-dimensions of the "quality of university teaching"), different methods proposed in the literature were taken into account. The following selection criteria were used: Criterion of the eigenvalues equal to or greater than 1 (Kaiser 1960), Parallel Analysis—PA (Horn 1965) and Minimum Average Partial—MAP (Velicer 1976). In all the factor analyses, the selection criteria identified four significant factors corresponding to the first four eigenvalues. Therefore, the initial solution and all the others show the four factors (CNVSU 2010, Sect. 1.3.5). The total inertia is explained satisfactorily: for both scales the variance accounted for is close to 60 %. In order to facilitate interpretation, we tried alternative solutions through the rotations of the factors. The most convincing solution is reached by performing a Varimax rotation. The items are grouped together following exactly the four sections of the standard questionnaire ("Organization of the course of study", "Organization of the teaching", "Teaching and study", "Infrastructure"). This last result is coherent with the preliminary item analysis conducted with the Cronbach's α. Finally, recalling that the classical item analysis had suggested significant associations between items of the different sections of the standard questionnaire (i.e. factors are correlated), here the more appropriate oblique rotations (Promax) using the polychoric correlation matrix was used. Table 2 shows the best separation of the items, obtained for UniBS with the 10-point scale.[4] Note that only the item "Supplementary activities are useful" is not quite coherent with its section in the questionnaire, and that items D14 and D15 belong to the factor of section three "Teaching and study".

The selection of variables to retain in the questionnaire can be made considering the results of factor analysis. The criterion used, shared by many authors to evaluate the importance of the variables is as follows: if the factor loadings are between 0.32 and 0.44 the contribution of the variable is negligible, between 0.45 and 0.54 is small, between 0.55 and 0.62 is good, between 0.63 and 0.70 is very good and 0.71 or more is excellent (Tabachnick and Fidell 2001).

Following this criterion, looking at the results of factor analysis we can define three groups of variables in relation to their importance. In the first group are the following important variables: D01, D02, D03 (except in one case), D04, D05, D07, D08 and D13, in the second we include variables that are mid-level D09, D10, D11, D12, D13 and D14, the third group is the only variable that D06.

[4]Some different results were obtained for UniSN and the 4-point scale (CNVSU 2010).

Table 2 Promax factor analysis for the 10-point scale of the standard questionnaire (UniBS data)

Item	Question	Factor 1	Factor 2	Factor 3	Factor 4
D01	Charged study acceptable	*0.886*	−0.092	−0.060	−0.027
D02	Overall organization acceptable	*0.868*	0.027	−0.069	−0.025
D03	Clarity of exam modality	0.073	*0.645*	0.038	−0.047
D04	Teaching timetable respected	−0.105	*0.825*	−0.151	0.005
D05	Teacher available for explanation	−0.029	*0.605*	0.169	−0.003
D06	Sufficient prior knowledge	0.192	−0.103	*0.374*	0.056
D07	Teacher stimulate the interest	−0.109	−0.082	*0.972*	−0.002
D08	Teacher exposes clearly	−0.071	0.008	*0.869*	−0.032
D09	Charged study proportional to the CFU	0.225	0.190	*0.350*	−0.066
D10	Appropriate teaching materials	0.035	0.287	*0.441*	−0.020
D11	Supplementary activities are useful	0.042	−0.034	0.203	*0.291*
D12	Appropriate classroom for lessons	0.144	0.205	−0.011	*0.341*
D13	Appropriate classroom for exercitations	−0.055	−0.033	−0.052	*0.929*
D14	Interest for the topics	0.082	0.023	0.607	−0.013
D15	Overall satisfaction	−0.014	0.070	0.854	−0.003

Now we have to verify the stability of parameters or the replicability of pattern/structure coefficients (loadings). So we have performed a nonparametric bootstrap factor analysis. As noted by (Thompson 1988), one of the difficulties of applying bootstrap procedure in factor analysis is the requirement that factors are presented in a common result space. Similar factors may be extracted across various sample, yet the factors may not occur in the same order as ranked by magnitude of the eigenvalues (factor II from one analysis may occur as factor III in another analysis). Hence it has been recommended that each alternative analytic run should be rotated to best fit a given "target matrix" derived from original factor run. A target matrix was created from the original Promax by placing a 1 (or −1) in the corresponding entry of the target matrix if a variable is connected to factor and a 0 on the remaining entries in the corresponding row (the Promax matrix was projected into same factor space as target matrix). Then we have drawn samples with replacement from original data matrix: in each resample, rows of the matrix data are randomly selected and each resample has exactly the same size as the original sample. The resampling was conducted 500 times and for each resample we have computed the polychoric correlation matrix and performed factor analysis. The Promax matrices were projected into same factor space as this target matrix. So we have computed the mean eigenvalues for each factor over 500 repeated resamples. We also compute the empirically estimated Standard Error (SE) of the estimates. In this manner we can determine if the mean of resampled eigenvalues is greater than 1 and if SE's of mean estimates near 1.0 are large or small.

At first, the bootstrap method is used to determine the number of factor to extract.[5] This approach suggested four eigenvalues greater than one, confirming the previous results. After determining the number of factors to extract we have to consider the parameters (loadings). The average bootstrap results, the estimated SE's and the ratio of the average bootstrap results to the estimated SE's can be used for either descriptive or inferential purposes. Small SE's indicate parameter stability across resampling. Therefore, if pattern coefficients are large in magnitude, the SE's tend to be smaller and the parameter estimates more stable. If the sample statistic is relatively equal to the mean bootstrap estimate and SE is small in relation to the mean bootstrap estimate, then the sample statistic can be thought of as stable. On the other hand, if the sample statistic and bootstrap estimate are not close or if the SE is large in relation to its mean bootstrap estimate, then bias may be reflected and caution is warranted when interpreting the sample data. The sample statistics and bootstrap estimates for variables D01 and D02 are close on factor I with relatively small SE's, for variables D03, D04 and D05 are close on factor II with relatively small SE's, for variables D06, D07, D08, D09 and D10 are close on factor III, but SE's for variables D06 and D09 are very large, for variables D11, D12 and D13 are close on factor IV but the SE for D11 is very large and, finally, the variables D14 and D15 are close on factor III with relatively small SE's.

The ratio of the mean bootstrap results and the SE that behaves like a t statistic is greater than 2.0 for each variable, deemed salient to given factors indicating that these coefficients are not zero. Therefore we can be confident that the results replicate over the resample and that the variables are correctly connected to the factors.

Finally, in order to evaluate the importance of items in the standard questionnaire with respect to the overall satisfaction for the university teaching, the logistic regression analysis has been used: the "Overall satisfaction" (item D15, with coding 0 and 1) was regressed onto the other 14 items. The final Table 3 displays the results obtained with the two different methods (factorial analysis—FA, and logistic regression using the forward selection approach—LR): the items that could be removed, because they contribute less than the others in explaining the overall satisfaction, are marked with X.

Rasch Analysis Results

In this paragraph we use the Rasch Analysis in the educational services evaluation context (King and Bond 2003; De Battisti et al. 2005; Bacci 2006; Pagani and Zanarotti 2010), to detect the appropriate ordinal response category structures among those considered. First of all, a preliminary data analysis has been performed in order to assess the effects of the 4- and 10-point response scales in disjoint (standard questionnaire and experimental questionnaire version B) and joint form

[5]Results of the bootstrap factor analysis are available on request to the authors.

Table 3 Results of factorial analysis (FA) and logistic regression (LR) for the items contribution in explaining the overall satisfaction (item marked X could be removed)

Method, Uni, scale	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12	D13	D14
FA not-rotated, UniBS, 4-point						X				X	X	X		X
FA rotated, UniBS, 10-point						X			X	X	X	X		
FA rotated, UniBS, 4-point						X			X	X	X	X		
FA not-rotated, UniSN, 4-point			X			X						X		X
FA rotated, UniSN, 4-point			X			X				X	X	X		
LR, UniBS, 4-point	X		X	X		X			X	X		X	X	
LR, UniBS, 10-point	X		X	X	X	X			X			X	X	
LR, UniSN, 4-point	X	X	X		X	X	X	X	X			X	X	X

(experimental questionnaire version A).[6] Figure 1 shows the percentage distributions of the students' responses on the 10-point scale with respect to the 4-point scale for the two versions of the experimental questionnaires.[7]

We observe that, for the experimental questionnaire version A distributions in Fig. 1a, the answers are not concentrated on few of the 10-point scale and are not uniformly distributed in the classes defined by the 4-point scale. In other terms, students have used for their responses all the range of the 10-point scale, differentiating their evaluation within the 4-point scale. We can see that, for the experimental questionnaire version B distributions in Fig. 1b, the answers on the 10-point scale are not constrained by the 4-point scale: students have used a more wider range of values than those bound by the 4-point scale. This is more evident in points 3 (13 % Definitely NO and 10 % More NO than yes), 6 (15 % More NO than yes and 16 % More YES than no) and 8 (31 % More YES than no and 21 % Definitely YES). This preliminary analysis shows that the 10-point scale allows students to express their assessment in more detail, and is not always entirely consistent with the 4-point scale.

The Rasch Analysis with the Rating Scale Model (Wright and Masters 1982) has been used to assess the properties of all the ordinal 4- and 10-point scales of

[6]For a sake of brevity, in this section only UniBS data analysis is presented, but with UniSN data we obtained roughly the same results (CNVSU 2010, Sect. 1.4).

[7]These percentage distributions for the standard questionnaire are roughly the same as the experimental questionnaire version B (CNVSU 2010, Sect. 1.3.4).

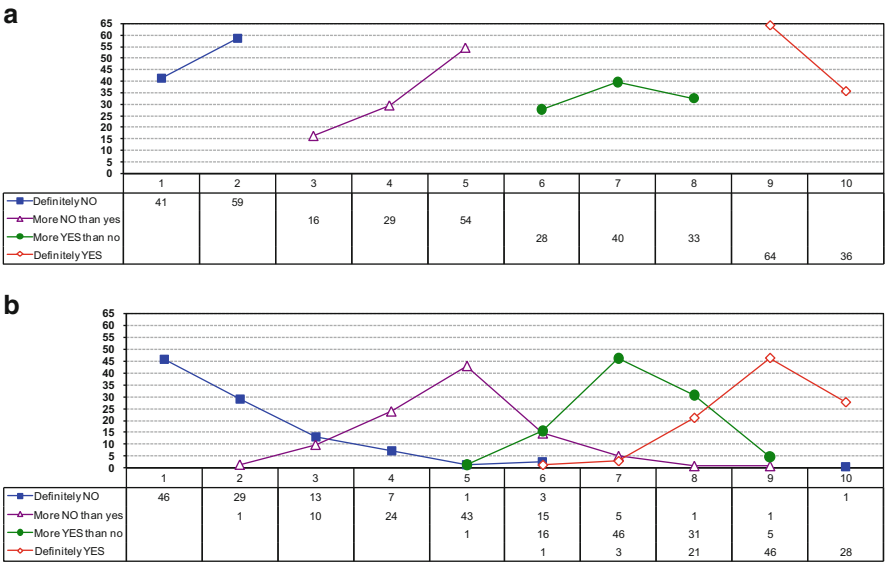

Fig. 1 Percentage distributions of the students' responses on the 10-point scale with respect to the 4-point scale for the two versions of the experimental questionnaires; (**a**) 4,234 responses for the experimental questionnaire version A (UniBS data); (**b**) 4,648 responses for the experimental questionnaire version B (UniBS data)

response used in the questionnaires. According to this model,[8] the probability that student i could give an answer x about the item j with $(c+1)$ ordered response categories is obtained as:

$$\pi_{ijx} = P\left(X_{ij} = x\right) = \frac{\exp \sum_{h=0}^{x} \left[\gamma_i - \left(\delta_j + \tau_h\right)\right]}{\sum_{k=0}^{c} \exp \sum_{h=0}^{k} \left[\gamma_i - \left(\delta_j + \tau_h\right)\right]} \qquad x = 0, 1, \ldots, c$$

where $\tau_0 \equiv 0$, so that $\exp \sum_{h=0}^{x} [\gamma_i - (\delta_j + \tau_h)] = 1$.

The probability π_{ijx} depends on the student attitude and item difficulty to be endorsed. The parameter γ_i identifies the "level of attitude" of student i, δ_j the mean difficulty to endorse item j and τ_h—the "threshold"—is the point of equal probability of categories $(h-1)$ and h. As goodness of fit statistic we use the *Rasch's Alpha* (RA) index, the raw *Score to Measure correlation* (SM) index and

[8] We have reasonably assumed that students perceive the items of the questionnaire to share the same rating (4 or 10 points) scale: in this case the Rating Scale Model is strongly to prefer with respect to others (for example the Partial Credit Model), as it is more parsimonious in the number of parameters, has higher estimate stability, and the communication of results is more easy (Linacre 2000).

Table 4 Threshold parameters for the 4- and 10-point scales of the standard and experimental questionnaires (UniBS data)

Questionnaire	4-Point scale			10-Point scale							
Standard	−1.35	−0.76	2.11	−0.48	−0.69	−0.90	−0.63	−0.26	−0.57	0.47	1.33	1.72
Experimental A	−1.88	−0.76	2.64	−0.96	−0.58	−0.82	−0.54	−0.45	−0.17	0.69	1.15	1.67
Experimental B	−1.56	−0.71	2.27	−0.47	−0.89	−0.90	−0.57	−0.39	−0.62	0.69	1.23	1.91

Table 5 Difficulty parameters for the 4- and 10-point scales of the standard and experimental questionnaires (UniBS data)

4-Point scale						10-Point scale					
Standard		Experimental A		Experimental B		Standard		Experimental A		Experimental B	
D06	1.15	D05	1.09	D05	0.82	D06	0.47	D05	0.49	D05	0.40
D01	0.86	D06	0.86	D01	0.75	D01	0.35	D06	0.37	D01	0.35
D02	0.66	D01	0.36	D08	0.54	D02	0.31	D01	0.19	D08	0.30
D07	0.48	D07	0.27	D06	0.37	D13	0.20	D07	0.12	D06	0.17
D12	0.45	D03	−0.13	D07	−0.16	D07	0.18	D03	−0.08	D07	−0.07
D13	0.35	D08	−0.53	D03	−0.38	D12	0.18	D08	−0.18	D03	−0.22
D08	0.20	D04	−0.63	D04	−0.78	D14	0.02	D04	−0.29	D04	−0.41
D14	0.15	D02	−1.30	D02	−1.17	D08	0.00	D02	−0.63	D02	−0.53
D09	−0.16					D11	−0.03				
D10	−0.27					D09	−0.09				
D11	−0.34					D10	−0.13				
D03	−0.79					D03	−0.33				
D05	−1.12					D05	−0.51				
D04	−1.62					D04	−0.63				

the *Explained Variance* (EV) index. The interpretation and the evaluation of the results for each item is based on three standard statistics used in the Rasch Analysis: *Difficulty*, *Infit* and *Ptmea*.

The Rasch Analysis allows us to place the items of a questionnaire on a continuum on the basis of their difficulty to be satisfied, thus defining for the respondents an evaluation measure on the same scale. This approach assumes the unidimensionality of the construct, which for our case roughly corresponds to the "quality of the university teaching".[9] Tables 4 and 5 show the obtained threshold and difficulty parameters for the 4- and 10-point scales of the standard and experimental questionnaires administered to the students of UniBS.

The model fit for UniBS data is rather good for the 15 items of the standard questionnaire (CNVSU 2010, Table 1.26), with all the Infit and Outfit statistics in the range from 0.6 to 1.4 as expected (Bond and Fox 2007, Table 12.6). Moreover, the correlations of the obtained Rasch measure with the sum score (0.90 for the

[9]We assumed that this "*principal dimension*" is defined by the sub-dimensions identified with the factor analysis in the previous paragraph.

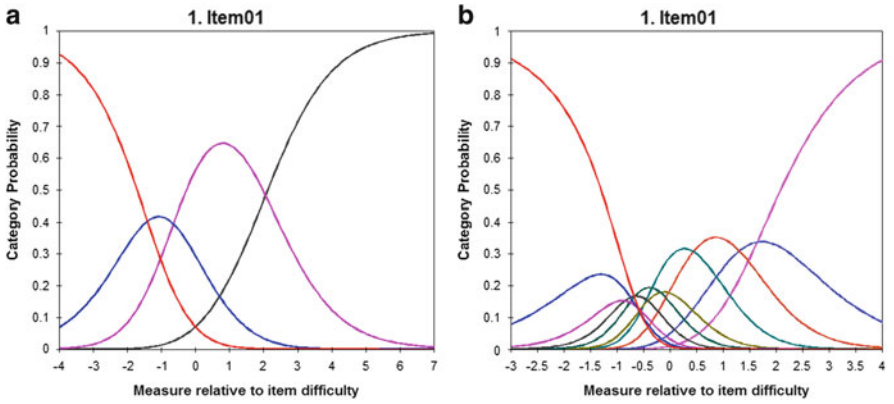

Fig. 2 Category probability curves for the standard questionnaire (UniBS data); (**a**) 4-point scale; (**b**) 10-point scale

4-point scale and 0.85 for 10-point scale) and with the overall satisfaction item (0.62 for the 4-point scale and 0.71 for 10-point scale) are quite higher.[10]

With the rating scale model the contribution of each item (question) to the measure construction can be evaluated. Two or more items positioned on the same point of the continuum (i.e. with the same difficulty to be endorsed) do not contribute substantially to the measure definition for the respondents. For the standard questionnaire, items with the same difficulties are for the 4-point scale D02–D07, D12–D13, D08–D14, D09–D10 and for the 10-point scale D07–D12–D13, D08–D14, D09–D10–D11.

Figure 2 shows the category probability curves for the standard questionnaire with 4- and 10-point scales.[11] For the 4-point scale in Fig. 2a, the shapes of these curves are coherent with the order of response categories (Definitely NO, More NO than yes, More YES than no, Definitely YES): higher probability is given sequentially to the four ordered categories, moving from the left to the right of the continuum. Note that the second category (More NO than yes) has a lower probability curve, which highlights a certain difficulty of discrimination, especially with regard to the first category.

For the 10-point scale in Fig. 2b the category probability curves tend to overlap for many votes (1, 2 and 3; 4 and 5; 6 and 7): the rating is not coherent with the order of the response categories moving from the left to the right of the continuum.

[10]We have used the correlation ratio for the 4-point scale and the linear correlation coefficient for the 10-point scale.

[11]As we used the parsimonious Rating scale model, the category probability curve is unique for all items of the same questionnaire.

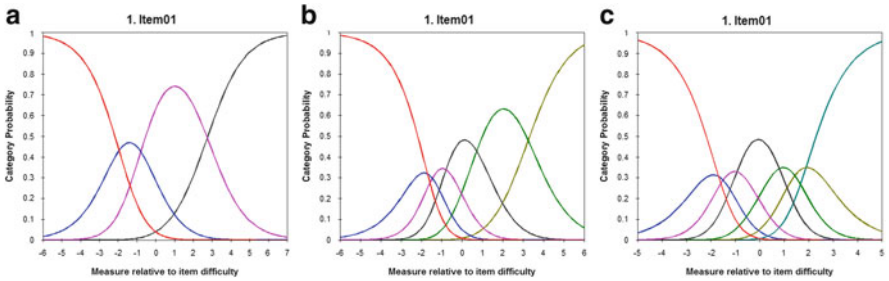

Fig. 3 Category probability curves for the standard questionnaire with new scales (UniBS data); (**a**) 4-category scale; (**b**) 6-category scale; (**c**) 7-category scale

The Rasch analysis has been conducted for the two versions of the experimental questionnaires with nine items. Obviously, fewer questions were redundant, but with similar scaling problems (see for details CNVSU 2010).

Previous results and further analyses have suggested to try to merge votes of the 10-point scale, obtaining the following 4-, 6- and 7-category scales[12]:

4-category scale: C1 = 1–2, C2 = 3–4–5, C3 = 6–7–8, C4 = 9–10

6-category scale: C1 = 1, C2 = 2–3, C3 = 4–5, C4 = 6–7, C5 = 8–9, C6 = 10

7-category scale: C1 = 1, C2 = 2–3, C3 = 4–5, C4 = 6–7, C5 = 8, C6 = 9, C7 = 10.

Figure 3 shows the category probability curves of these new category scales for the standard questionnaire. For the 4-point scale in Fig. 3a the shapes of the curves are very similar to those of the ordered response categories in Fig. 2a: merging the 10-point scale in a 4-point scale we find the same pattern of the probability curves, except for the two extreme ones, that are more symmetrical with respect to zero.

For the 6-point scale in Fig. 3b the category probability curves show an overlapping for categories C1–C2 and C3–C4. For the 7-point scale in Fig. 3c the shapes of the probability curves are symmetrical with respect to the central category C4, with an overlapping especially for C1–C2 and C6–C7.

Concluding Remarks

In this study, using data from a recent survey, factor pattern matrices of the questionnaire of university teaching evaluation were examined for simple structure and interpretability. As shown in Table 2, the four-factor solution resulted in a clear

[12]Note that the 4- and 6-category scales use symmetric merging of the 10 votes with respect to negative (less than 5) and positive (more than 5) evaluations.

Furthermore, with these three new scales we could use the following

4-point scale: 1.5, 4, 7, 9.5

6-point scale: 1, 2.5, 4.5, 6.5, 8.5, 10

7-point scale: 1, 2.5, 4.5, 6.5, 8, 9, 10.

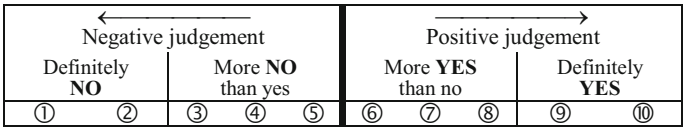

Fig. 4 The proposed scale for the evaluation of university teaching

and interpretable structure, Factor 1 contained two items with salient factor pertinent to "Organization of the course of study", Factor 2 contained three items pertinent to "Organization of the teaching", Factor 3 contained seven items pertinent to "Teaching and study" (but item 6 is weak) and Factor 4 contained 3 items pertinent to "Infrastructure" (but items 11 and 12 are weak).

The scaling analysis conducted with the Rating scale model is not conclusive, but it shows the greater flexibility offered by the 10-point scale: its votes can be merged in a 4-point scale preserving the comparisons with previous surveys, and more informative statistical analysis could be conducted for the future surveys. For this reason, we propose to consider for the responses of the evaluation of university teaching the scale in Fig. 4.

As a final product of this research project, a new short version of questionnaire has been proposed (CNVSU 2010, Chap. 3).

References

Bond, T.G., Fox, C.M.: Applying the Rasch Model, 2nd edn. LEA, London (2007)

CNVSU: Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti, Rapporto di Ricerca, RdR 1/00 (2000)

CNVSU: Progettazione, implementazione e validazione di un questionario per la valutazione della didattica erogata a studenti universitari. Rapporto di Ricerca, RdR 2/10. www.cnvsu.it/_library/downloadfile.asp?id=11775 (2010)

Bacci, S.: I modelli di Rasch nella valutazione della didattica universitaria. Statistica Applicata **18**(1), 5–49 (2006)

De Battisti, F., Nicolini, G., Salini, S.: The Rasch model to measure the service quality. ICFAI J. Serv. Mark. **III**(3), 58–80 (2005)

Harman, H.H.: Modern Factor Analysis. University of Chicago Press, Chicago (1960)

Horn, J.L.: A rationale and test for the number of factors in factor analysis. Psychometrika **30**, 179–185 (1965)

Jöreskog, K.G.: Factor Analysis by Minres. Scientific Software, Chicago (2003). www.ssicentral.com/lisrel/techdocs/minres.pdf

Kaiser, H.F.: The application of electronic computers to factor analysis. Educ. Psychol. Meas. **20**, 141–151 (1960)

King, J.A., Bond, T.G.: Measuring client satisfaction with public education I: meeting competing demands in establishing state-wide benchmarks. J. Appl. Meas. **4**(2), 111–123 (2003)

Linacre, J.M.: Comparing "partial credit" and "rating scale" models. Rasch Meas. Trans. **14**(3), 768 (2000)

Nunnally, J.C., Bernstein, I.H.: Psychometric Theory. McGraw-Hill, New York (1994)

Pagani, L., Zanarotti, C.: Some uses of Rasch models parameters in customer satisfaction data analysis. Qual. Technol. Quant. Manag. **7**(1), 83–95 (2010)

Tabachnick, B.G., Fidell, L.S.: Using Multivariate Statistics, 4th edn. Allyn & Bacon, Needham Heights (2001)
Thompson, B.: Program FACSTRAP A program that computes bootstrap estimates of factor structure. Educ. Psychol. Meas. **61**, 681–686 (1988)
Velicer, W.F.: Determining the number of components from the matrix of partial correlations. Psychometrika **41**, 321–327 (1976)
Wright, B.D., Masters, G.N.: Rating Scale Analysis. MESA, Chicago (1982)

A Family of Indices for Teaching Evaluation: Experiences in Italian Universities

Donata Marasini and Piero Quatto

Abstract

In order to analyze the student ratings of University teaching, several indices summarize the relative frequencies of positive and negative responses in a single numerical value. Focusing on linear functions of response frequencies, the paper studies some interesting families of indices for the measurement of student satisfaction. Special attention is paid to relationships between these families and to a particular family that arises in a natural way.

Introduction

Italian Universities are required to carry out a survey on student satisfaction and teaching facilities. CNVSU (National Evaluation Committee of the University System) annually requests Evaluation Commissions (ECs) to provide a summary by means of the percentages of positive and negative ratings. In particular CNVSU has provided a questionnaire outline that Universities have to follow in terms of content but not necessarily in terms of form, meaning that any CNVSU item should be present in the questionnaire but the answer can be articulated in different ways. For example, such a questionnaire may be consistent with the requirements of CNVSU involving four response categories: decidedly no (*DN*), more no than yes (*MN*), more yes than no (*MY*), decidedly yes (*DY*); otherwise, the questionnaire can be structured on a far richer and more sensitive scale. Having chosen an ordinal scale, ECs summarize student ratings by means of suitable indices depending on absolute or relative frequencies of student responses (CNVSU 2002; Cocuzzoli et al. 2007;

D. Marasini (✉) • P. Quatto
Dipartimento di Statistica, Università degli Studi di Milano – Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milan, Italy
e-mail: donata.marasini@unimib.it; piero.quatto@unimib.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_25,

Gruppo Valmon; Nuclei 2010). Such an index can be either qualitative, as the median, or quantitative, as a weighted mean of the frequencies. In both cases we obtain a descriptive index as opposed to a model-based index, according to Rampichini et al. (2004).

This paper focuses on weighted mean of the relative frequencies corresponding to responses *DN*, *MN*, *MY*, and *DY*, since the 2010 CNVSU survey referring to the academic year 2008–09 showed that about 60 % of the Universities (excluding telematics Universities) also uses such a weighted mean for the assessment of student satisfaction. Diversity of weights adopted by several Universities has made appropriate a formalization of the various proposals. In this way, we were able to show that some proposals are different only in appearance.

Specifically, Sect. 2 introduces two families, denoted by IS_w and IS_k, grouping many of the indices used by Italian Universities. In particular, we focus on the subfamily IS_{w*} of IS_w that has certain characteristics and we verify the equality of the corresponding standardized families. Then we show that there is a linear relationship between elements belonging to IS_{w*} and elements belonging to IS_k. In Sect. 3 we introduce three other indices which are elements of the family IS_w only under particular conditions, but does not belong to the subfamily IS_k. This family involves a four-point scale, but there are Universities that use larger scales. In this regard, Sect. 4 considers the 1-to-10 scale, as presented by experimental research which showed some uncertainties confirming the validity of four-point scale. Then we provide a few indications for the choice of one among the indices belonging to IS_k. Finally, Sect. 5 extends the study to the so-called multidimensional indices.

The Families IS_w and IS_k

With reference to a specific item of the questionnaire (such as "Overall, are you satisfied with this course?), let p_1, p_2, p_3, and p_4 be relative frequencies corresponding, respectively, to responses *DN*, *MN*, *MY*, and *DY*.

Student satisfaction indices employed by many Universities belong to the following family

$$IS_w = \sum_{j=1}^{4} w_j \, p_j \,, \tag{1}$$

where the components w_j of the vector $w = (w_1, w_2, w_3, w_4)$ are the weights assigned to the relative frequencies. In this context, the weights are considered as predefined constants, such that $w_1 \leq w_2 \leq w_3 \leq w_4$.

For instance, if $w = (1, 2, 3, 4)$ or $w = (2, 5, 7, 10)$, then (1) represents the weighted mean hereinafter denoted by $IS_{(1,2,3,4)}$ or by $IS_{(2,5,7,10)}$, respectively; the former (denoted by the acronym SDI) was suggested by Cerchiello et al. (2010) and the latter by Chiandotto and Gola (2000). They are used by many Italian Universities: e.g., $IS_{(1,2,3,4)}$ is employed by the University of Genoa, Verona, and

Milan Polytechnic; $IS_{(2,5,7,10)}$ is used by the eight Universities participating in the Valmon group (including the University of Florence, founder of the group). In the case of $w = (0, 0, 1, 1)$, we obtain the index adopted by CNVSU $IS_{(0,0,1,1)}$ that provides the percentage of positive ratings.

In general, if $w = (-1, -k, k, 1)$ with $0 \leq k \leq 1$, then (1) provides the subfamily of IS_w denoted by IS_k.

Among the indices included in IS_w, those satisfying the equality

$$w_4 - w_3 = w_2 - w_1 \tag{2}$$

with

$$w_1 < w_2 < w_3 < w_4 \tag{3}$$

seem to be of particular interest, due to the fact that (2) represents a condition of symmetry between the *DN* and *DY* values whatever the two central categories. So, we define the subfamily IS_{w*} of IS_w consisting of the indices satisfying (2) and (3). Thus, we have

$$IS_k \subset IS_{w*} \tag{4}$$

because the fact that the elements of IS_k with $0 < k < 1$ satisfy (2).

Besides the families of absolute indices, we consider the corresponding families of standardized indices. In particular, with regard to (1), we have

$$\tilde{IS_w} = \frac{\left(\sum_{j=1}^{4} w_j\, p_j\right) - w_1}{w_4 - w_1} = \sum_{j=1}^{4} \frac{\left(w_j - w_1\right) p_j}{w_4 - w_1} = \sum_{j=1}^{4} z_j\, p_j,$$

where

$$z_1 = 0 \quad z_2 = \frac{w_2 - w_1}{w_4 - w_1} \quad z_3 = \frac{w_3 - w_1}{w_4 - w_1} \quad z_4 = 1 \tag{5}$$

and with regard to IS_k we obtain

$$\tilde{IS_k} = \frac{IS_k + 1}{2} = \sum_{j=1}^{4} \frac{\left(w_j + 1\right) p_j}{2} = \sum_{j=1}^{4} z_j\, p_j, \tag{6}$$

where

$$z_1 = 0 \quad z_2 = \frac{1 - k}{2} \quad z_3 = \frac{1 + k}{2} \quad z_4 = 1. \tag{7}$$

With respect to the standardized version $I\tilde{S}_{w*}$, if we define

$$k = \frac{w_3 - w_2}{w_4 - w_1} \tag{8}$$

the coefficients z_2 and z_3 appearing in (7) are equal to their counterparts given in (5). Conversely, equating the coefficients z_2 and z_3 with those of (7), we obtain (2) and (8). Hence, it follows that $I\tilde{S}_{w*} = I\tilde{S}_k$. Recalling (6), we have $IS_k = 2I\tilde{S}_k - 1$ and this equality can be rewritten as

$$IS_k = 2I\tilde{S}_{w*} - 1 = 2\frac{IS_{w*} - w_1}{w_4 - w_1} - 1 = \frac{2}{w_4 - w_1} IS_{w*} - \frac{w_1 + w_4}{w_4 - w_1}. \tag{9}$$

Thus, (9) expresses the linear relationship between an element of IS_{w*} and an element of IS_k. So, with respect to the two indices $IS_{(1,2,3,4)}$ and $IS_{(2,5,7,10)}$ belonging to IS_{w*}, the linear relationships are $IS_{1/3} = \frac{2}{3} IS_{(1,2,3,4)} - \frac{5}{3}$ and $IS_{1/4} = \frac{1}{4} IS_{(2,5,7,10)} - \frac{6}{4}$, where $IS_{1/3}$ is adopted by University of Insubria, while $IS_{(1,4,7,10)}$, related to $IS_{1/3}$ through the relationship $IS_{1/3} = \frac{2}{9} IS_{(1,4,7,10)} - \frac{11}{9}$, is adopted by the University of Bologna. It is important to note that (9), being a strictly increasing function, maintains the ranking; this means that if course is ranked r using IS_{w*} the same value remains if the corresponding index IS_k is used. From the two conditions $IS_k \subset IS_{w*}$ and $I\tilde{S}_k = I\tilde{S}_{w*}$ it follows that the relationship between IS_k and IS_{w*} is not bijective, in the sense that there may be more indices belonging to the family IS_{w*} that give rise to the same index of the family IS_k. It can be proved that all and only the indices IS_{w*} with the same image IS_k have coefficients given by

$$w^* = \left(w_1, \frac{w_1 + w_4}{2} - \frac{k}{2}(w_4 - w_1), \frac{w_1 + w_4}{2} + \frac{k}{2}(w_4 - w_1), w_4 \right). \tag{10}$$

In this context, two indices with coefficients (10) are defined equivalent if they preserve the same ordering. Thus, indices IS_{w*} give rise to the equivalence class $\{IS_k\}$. In other words, $\{IS_k\}$ contains all the indices IS_{w*} that provide the same ordering of IS_k. Another index used in some Universities, for example Genoa, IULM, Pavia, and Biomedical Campus of Rome, is given by $k = 1/2$ and the corresponding equivalence class $\{IS_{1/2}\}$ consists of all and only the indices IS_{w*} with

$$w^* = \left(w_1, \frac{3w_1 + w_4}{4}, \frac{w_1 + 3w_4}{4}, w_4 \right).$$

For instance, $IS_{(2,4,8,10)}$ (employed by the University of Naples Parthenope) is an element of this class.

Other Indices

This section deals with three indices that, under suitable restrictions, belong to the family IS_w, but do not belong to any of the subfamilies taken into account.

With respect to the first index, if we define $w_j = \sum_{i=1}^{j} p_i' - \frac{1}{2} p_j'$, where p_i' represents the relative frequency to the response i in a University considered as a reference, ($j = 1, 2, 3, 4$), we obtain the well-known Ridit (Bross 1958; Agresti 1984), which takes the form

$$IS_{RI} = \sum_{j=1}^{4} \left(\sum_{i=1}^{j} p_i' - \frac{1}{2} p_j' \right) p_j .$$

If the frequencies p_i' are known, such an index belongs to IS_w. Condition (2) is generally not met, therefore IS_{RI} does not belong to the family IS_{w*} or to its subfamily IS_k and the ratio (8) does not hold.

The University of Milano-Bicocca some time ago adopted the following linear transformation of IS_{RI}

$$IS_{AG} = 2IS_{RI} - 1$$

proposed by Agresti (1981).

The third index (Civardi et al. 2006), also employed long ago by the University of Milan-Bicocca, can be written as (1) with

$$w_1 = -\frac{1}{1+h}\left(1 + \frac{h}{p_1 + p_2}\right), \quad w_2 = -\frac{1}{1+h},$$
$$w_3 = \frac{1}{1+h}, \quad w_4 = \frac{1}{1+h}\left(1 + \frac{h}{p_3 + p_4}\right)$$

($p_1 + p_2 > 0, p_3 + p_4 > 0, 0 \le h \le 1$). Since the weights depend on the corresponding frequencies p_j , such an index does not belong to the family IS_w, except when $h = 0$. In this case we have $w = (-1, -1, 1, 1)$ and the index represents the difference between positive and negative ratings.

Considerations on the Choice of a Suitable Index

The question is now the choice of an index among those considered above.

First of all, it is advisable to choose one of the two subfamilies IS_k or IS_{w*}. The choice is almost immediate because, given the inclusion (4) and the relationship (9) which maintains the ranking, it is possible to focus on IS_k. We can observe that the family (1) could be generalized to the case of a h point scale as follows:

$$IS_w = \sum_{j=1}^{h} w_j \, p_j \, .$$

with $w = (w_1,....,w_h)$ and $w_1 < \cdots < w_h$.

On the other hand, IS_k involves a four-point scale. In the following, the use of such a scale is motivated by the survey carried out by a research group of University of Brescia and Sannio presented in the report "Design, implementation and validation of a student questionnaire to assess University courses" published by CNVSU in 2010. With specific reference to the survey conducted in Brescia, each of the students involved filled in the standard questionnaire consisting of 15 items. Apart from the traditional four responses (*DN*, *MN*, *MY*, *DY*), students were asked for a score from 1 to 10 regardless of the previous answers. Each student, however, also filled in another experimental questionnaire with nine items (plus an extra one asking to express a preference between the standard and experimental questionnaire). For each item, the response *DN* was scored 1 and 2; the response *MN* was scored 3–5; the response *MY* was scored 6–8; and the response *DY* was scored 9 and 10. The student only indicated his or her score, from which one of the four responses could be retrieved; in other words if the score was 6, the response was *MY*. When measuring the consistency of the students' responses in the case of the standard questionnaire, the following occurred: among the students who answered *DN*, as many as 46 % gave a score of at least 3, even though the mode ranked 1 with 28 %. Among those who answered *DY*, 38 % attributed a score between 6 and 8, while the mode ranked 9 with 38 %. This shows that, although the mode values are 1 for *DN* and 9 for *DY*, high percentages attributed scores higher than 2 for *DN* and lower than 9 for *DY*. In other words, those who responded *DN* were not so sure in attributing an extremely low score, and those who responded *DY* were not so sure in attributing an extremely high score, which shows a sort of dangerous inconsistency between judgement and score. In the case of the experimental questionnaire, on the contrary, the mode score corresponding to *DN* was 2 with 59 %, the mode score corresponding to *MN* was 5 with 54 %, the mode score corresponding to *MY* was 7 with 40 %, and the mode score corresponding to *DY* was 9 with 64 %; with this version, the inconsistency of responses is in some way reduced by the joint scale balance. Excluding the ten-point scale, the choice is between the four points or the four points together with ten scores as recommended by the research group. The four-point scale alone appears to us to be more incisive compared with the other one because it prevents uncertainty among the various possibilities.

Having chosen the family IS_k, we now have to choose an element in it.

A possible criterion states that the difference between the two categories *DN* and *MN* should be same as between *MN* and *MY*. Such a criterion implies that a student's choice between, for example, a totally negative response as opposed to a partly negative one, is the same as the choice between a partly negative and a partly positive response. Given that the difference in absolute value between the first pair is $(k-1)$ and between the second is $2k$, homogeneity is guaranteed by $k = 1/3$. The value $k = 1/3$ identifies the equivalence class $\{IS_{1/3}\}$, that contains $IS_{(1,2,3,4)}$.

On the other hand, if the choice between *DN* and *MN*, and similarly between *MY* and *DY*, can be considered less problematic than the choice between *MN* and *MY*, then the results are $k > 1/3$ and $k < 1/3$ in the opposite case. In this way, we need to choose a value in each case.

In this regard, let X be a r.v. (random variable) which takes the four values w_1, w_2, w_3, w_4 with probability p_1, p_2, p_3, p_4. So, under condition (3), the equality $P(X = w_j) = p_j$ can be seen as the probability that one student, chosen at random, if asked to evaluate a course, will assign the value w_j.

Let Y be another r.v. which assumes the same values with probability $\pi_1, \pi_2, \pi_3, \pi_4$ taken as comparison variable with X.

Hence, $P(X \geq Y)$ represents the probability that one student chosen at random is at least as satisfied as another student taken as comparison.

Moreover, let

$$P\,(X \geq Y) - P\,(X \leq Y) \tag{11}$$

be the difference which, depending on whether it is higher, lower or equal to zero, shows a greater, smaller, or equal probability of satisfaction on the part of the first student compared to the second. Hence, we obtain

$$\begin{aligned} P\,(X \geq Y) - P\,(X \leq Y) &= \sum_{j=1}^{4} \left[P\left(Y \leq w_j\right) - P\left(Y \geq w_j\right)\right] p_j \\ &= (\pi_1 - 1)\, p_1 + (2\pi_1 + \pi_2 - 1)\, p_2 + (2\pi_1 + 2\pi_2 + \pi_3 - 1)\, p_3 + p_4 \end{aligned} \tag{12}$$

and (12) coincides with IS_k if, and only if, (a) the r.v. Y is characterized by the probability $\pi_1 = 0, \pi_2 = 1/2, \pi_3 = 1/2, \pi_4 = 0$ and (b) it turns out $k = 1/2$. Therefore (11) coincides with $IS_{1/2}$ and represents a difference in satisfaction between one student chosen at random and a situation which excludes the extreme situations *DN* and *DY* and expresses the maximum uncertainty between the two intermediate responses *MN* and *MY*. Thus, if we assume the hypothesis $k > 1/3$, then the previous argument leads to choose $k = 1/2$.

On the other hand, if we assume $k < 1/3$, no criterion seems to validate a specific value of k and we may only suggest to use $k = 1/4$, since $IS_{1/4}$ is the image of $IS_{(2,5,7,10)}$ which is the index most used in the Italian experience.

Multidimensional Indices

Regardless of the choice of k, the corresponding index IS_k is simple, in the sense that it concerns only one item of a specific course; it can be made multidimensional, by extending it to all the items in one course (Rampichini et al. 2004), and to all the courses of interest.

In this regard, we can introduce a suitable dispersion index associated with IS_k. Following Leti (1983), an ordinal dispersion index measures the dispersion with respect to the polarization given by $p_1 = p_4 = 1/2$.

It can be proved that the variance represents an ordinal dispersion measure and that

$$D_k = \left(\frac{2}{w_4 - w_1}\right)^2 D_{w*}, \tag{13}$$

where D_{w*} and D_k are the variances associated with IS_{w*} and, respectively, with the corresponding IS_k given by (9). In particular, we obtain $D_k = \tilde{D}_k = \tilde{D}_{w*}$ for standardized indices. Therefore, the equivalence class $\{IS_k\}$ is associated with the class $\{D_k\}$ whose elements satisfy (13).

In the Italian experience, items included in the questionnaire administered to students are divided into five groups: (1) organization of the course of study, (2) organization of each course, (3) teaching and learning, (4) infrastructure, (5) interest and satisfaction. For example, with respect to group (3), let Q be the number of items included in the questionnaire, C be the number of courses in a Course of study, and R the number of Courses of study in one Faculty. Furthermore, let $IS_{k(rcq)}$ and $D_{k(rcq)}$ be an index of IS_k and, respectively, the associated ordinal dispersion index, calculated on the item q ($q = 1, \ldots, Q$) and the course c ($c = 1, \ldots, C$) of the Course of study r ($r = 1, \ldots, R$). Let n_{rc} be the number of questionnaires related to the course c of the Course of study r. So, a weighted index is given by

$$IS_{k(rc)} = \frac{\sum_{q=1}^{Q} \left(1 - D_{k(rcq)}\right) IS_{k(rcq)}}{\sum_{q=1}^{Q} \left(1 - D_{k(rcq)}\right)} \tag{14}$$

(Russo 2002) or by the standardized version

$$\tilde{IS}_{k(rc)} = \frac{\sum_{q=1}^{Q} \left(1 - D_{k(rcq)}\right) \tilde{IS}_{k(rcq)}}{\sum_{q=1}^{Q} \left(1 - D_{k(rcq)}\right)}.$$

Finally, the index given by

$$\tilde{IS}_{k(r)} = \frac{\sum_{c=1}^{C} n_{rc} \tilde{IS}_{k(rc)}}{\sum_{c=1}^{C} n_{rc}}$$

leads to the multidimensional index

$$\tilde{IS}_k = \frac{\sum_{r=1}^{R} \omega_r \tilde{IS}_{k(r)}}{\sum_{r=1}^{R} \omega_r},$$

where ω_r is a weight assigned to the Course of study r depending on peculiarities of the Course. For example, ω_r may be the number of student enrolled in the Course or the total number of questionnaires collected $n_r = \sum_{c=1}^{C} n_{rc}$.

References

Agresti, A.: Measures of nominal–ordinal association. J. Am. Stat. Assoc. **76**(375), 524–529 (1981)

Agresti, A.: Analysis of Ordinal Categorical Data. Wiley, New York (1984)

Bross, I.D.J.: How to use Ridit analysis. Biometrics **14**(1), 18–38 (1958)

Cerchiello, P., Dequarti, E., Giudici, P., Magni, C.: Scorecard models to evaluate perceived quality of academic teaching. Stat. Appl. **2**, 145–156 (2010)

Chiandotto, B., Gola, M.: Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti, Comitato Nazionale per la Valutazione del Sistema Universitario, RdR1/100. www.cnvsu.it (2000)

Civardi, M., Crocetta, C., Zavarrone, E.: Summary indicators of opinion expressed by the users of a given service. Statistica **LXVI**(4), 373–388 (2006)

CNVSU: Proposta di un insieme minimo di domande per la valutazione della didattica da parte degli studenti frequentanti, Doc. 9/02. www.cnvsu.it (2002)

CNVSU: Progettazione, implementazione e validazione di un questionario per la valutazione della didattica erogata a studenti universitari, RdR 2/10. www.cnvsu.it (2010)

Cocuzzoli, P., Ingrassia, S., Costanzo, G.D., Mazza, A.: Indicatori statistici per la valutazione della soddisfazione didattica universitaria, Rivista di Economia e Statistica del Territorio, n. 3 (2007)

Gruppo Valmon. https://valmon.ds.unifi.it/sisvaldidat/index.php

Leti, G.: Statistica descrittiva. il Mulino, Bologna (1983)

Nuclei http://nuclei.miur.it/sommario/ (2010)

Rampichini, C., Grilli, L., Petrucci, A.: Analysis of university course evaluations: from descriptive measures to multilevel models. Stat. Methods Appl. **13**, 357–373 (2004)

Russo, M.A.: Indicatori statistici per la valutazione della qualità della didattica universitaria: una proposta metodologica. Statistica **LXII**(3), 431–452 (2002)

The Unity of Italy from the Point of View of Student Performances: Evidences from PISA 2009

Mariagiulia Matteucci and Marilena Pillati

Abstract

This paper investigates Italian student performances based on the 2009 edition of the Programme for International Student Assessment (PISA), a survey conducted by Organization for Economic Cooperation and Development (OECD) in order to assess skills of 15-year-olds in schools with respect to reading, mathematical and scientific literacy. In particular, student outcomes are compared among different Italian regions, taking into account socio-economic background and school membership of students. The results show that, despite the existence of a unified educational system in Italy, regional differences are evident. However, taking the nested structure of data into account, it can be shown that the most part of the variability in student performances can be explained by differences at school level.

Introduction

The anniversary of 150 years from the Unity of Italy is an occasion for taking stock of the Italian situation by different points of view. In particular, the study of territorial differences, continuously pointed out at economic, social and cultural level, assumes a particular importance.

In the educational field, performance gaps among the different areas of the country are observed regularly. National assessments, conducted by the National Evaluation Institute for the School System (INVALSI) at different school levels, reveal that geographical differences are evident with respect to Italian language and mathematics both in primary and lower secondary school (INVALSI 2010a, b).

M. Matteucci (✉) • M. Pillati
Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy
e-mail: m.matteucci@unibo.it; marilena.pillati@unibo.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_26,

In particular, students in the North of Italy usually outperform students in the Centre, and the most worrying deficiencies are observed for students living in the South of the country. Moreover, results of Northern regions are rather similar while the variability in performances of Southern regions is considerable, both within and between regions. These results are generally confirmed also by international surveys. Territorial gaps in learning and competencies mean not only a difference in the effectiveness but especially in the equity of the scholastic system, producing as a consequence economic and social costs.

The study of student performances in secondary education should be conducted by taking into account the different types of upper secondary schools. In fact, three main types of school are present (academic, technical and vocational) which involve a different rate of students in different territorial areas. The relative majority of students choose an academic track (liceo), especially in the Centre and in the Southern part of the country. On the other hand, the North-East is characterized by highest rate of students in technical or vocational schools.

The aim of this work is to understand which picture of the territorial uniformity emerges from the last international student assessment conducted in Italy: the OECD PISA 2009. The question then is: is Italy a unified country from the point of view of student performances? Also, does the scholastic system show the same limits and potentialities in all geographical areas? These issues are explored also taking into account the nested structure of data in order to understand how much of the variability in the performances can be imputed to geographical differences and to the school effect.

The paper is organized as follows. In Sect. 2, a short description of the OECD PISA survey is given, with details on the Italian case, while in Sect. 3 the main results on the Italian student performance are presented. Sect. 4 is devoted to the multilevel analysis of data and Sect. 5 addresses concluding remarks.

Features of the OECD PISA Survey

The Programme for International Student Assessment (PISA) is a standardized survey conducted by the Organization for Economic Cooperation and Development (OECD) and developed by participating economies (OECD 2010). The main aim of PISA is to assess reading, mathematical and scientific literacy of 15-year-olds in schools at international level. In particular, students near the end of compulsory education are evaluated not merely in terms of curricular contents but taking into account important knowledge and skills needed for full participation in society in adult life. To this end, students are expected to use their knowledge and capabilities in facing a wide range of texts and problems, not necessarily belonging to their scholastic or familiar experience (OECD 2009b). The idea underlying PISA is that the literacy level could be an indicator of the social capital and a predictor of socio-economic welfare of single individuals and countries. For this reason, the survey is

assuming an increasing importance for comparative purposes at international level and longitudinal perspective.

The survey started in 2000 and it is renewed every three years. For each edition, all three literacy domains are assessed but only one is privileged (reading in 2000 and 2009, mathematics in 2003 and science in 2006). Furthermore, student engagements and strategies are investigated, and information about schools, parents and socio-economic background are collected as well. The available information about students allow not only to evaluate the performances related to the three different domains, but also to study the relationship between the performances and the characteristics of students, schools and the whole system.

By using data on item responses, standardized literacy scales are constructed for each domain having a mean value of 500 and a standard deviation of 100. The methodology underlying this process is *item response theory* (Lord and Novick 1968; Hambleton and Swaminathan 1985). In particular, the mixed coefficients multinomial logit model (Adams et al. 1997) is used in order to set item difficulty and student ability on the same scale. The computation of the student individual score involves the concept of plausible values, which are simply random draws from the posterior distribution of ability. For each student, five plausible values are reported, which should be taken into account when computing score statistics. As a second step, literacy scales are divided into levels, i.e. proficiency levels, representing increasing item difficulties and student abilities. Proficiency levels can be used to classify students not only according to their score, but also describing what they can do in practice within each ability domain.

In Italy, the edition of 2009 was characterized by the presence of a representative sample for all regions and the two autonomous provinces of Trento and Bolzano. Even if Italy has a single scholastic system, the availability of regional data is motivated by the decentralization process sanctioned by the 2001 constitutional reformation and it represents an important source of information for studying the school gaps in different territorial areas of the country.

The PISA sample is a two-stage stratified sample, where the first-stage sampling units consist of individual schools having PISA eligible students (sampled with probabilities proportional to the school dimension) and the second-stage sampling units are students within sampled schools. For each school, a sample of 35 students was selected. The stratification design was planned so that reliable estimates could be obtained not only at national level, but also by school type and by regions. The final Italian sample consisted of 1,097 upper and lower secondary schools (only 1.5 % of lower secondary) and 30,905 students. Besides the regional stratification, a further territorial aggregation can be considered in Italy, dividing the country into five main geographical macro-areas: North-West (Piemonte, Lombardia, Liguria and Valle d'Aosta), North-East (Veneto, Friuli Venezia Giulia, Trento, Bolzano and Emilia-Romagna), Centre (Toscana, Lazio, Umbria and Marche), South (Abruzzo, Molise, Campania and Puglia) and South-Islands (Calabria, Basilicata, Sicilia and Sardegna).

Main Results of PISA 2009 in Italy

The most immediate and evident result from the last edition of the OECD PISA is that the overall performance of Italian students is statistically significantly lower than the OECD average in all the three domains of reading, mathematics and science. However, looking at the results within the single geographical areas and regions, shown in Table 1, the situation appears rather different.

In the reading scale, students of North-West and North-East obtain a mean score significantly higher than the Italian average and the OECD average, while students in the South and South-Islands achieved the worst results, significantly lower than both the Italian and the OECD mean score. The Centre of Italy obtains intermediate performances, with a mean score close to the Italian mean but statistically lower than the international mean. Exactly the same conclusions can be drawn looking at the results for the mathematical and scientific literacy scales. At a regional level, the top performing regions are Friuli Venezia Giulia, Lombardia, Veneto and Trento for all domains but all regions in Northern Italy obtain rather satisfying scores, at least close to the Italian and the OECD mean. When looking at the Southern part of the country, the results are significantly below the Italian and the OECD mean, with the exceptions of Abruzzo and especially of Puglia.

With respect to the distribution of students among the proficiency levels, what should be highlighted is the rate of low and high performers, defined by OECD 2011. Low performers are those students who do not attain the PISA baseline proficiency Level 2 in reading, at which the student is asked to determine the main idea of a text, understand relationships or infer meaning when the information is not prominent. High performers are those students who attain proficiency Level 5 or above, at which students must have a full and detailed understanding of a text whose content or form is unfamiliar. The highest percentages of top performers are present in Northern regions, where Emilia-Romagna, Friuli Venezia Giulia, Lombardia, Valle d'Aosta and Trento exceed 8.5 %. On the other hand, Calabria, Campania, Molise and Sicilia show very low percentages of top performers (less than 3 %). On the other hand, the number of low performers is worrying in the Southern regions and particularly for Calabria, Campania and Sicilia with more of the 30 % of cases. These regions show the most critical situation with respect to both the lack of excellent students and the presence of students who do not attain the baseline proficiency. On the other hand, the smallest percentages of low performers are present in Friuli Venezia Giulia, Lombardia, Valle d'Aosta, Veneto and Trento (less than 15 %).

By conditioning for school type, the ranking of the Italian macro-areas remains unchanged. In fact, the Northern areas got the highest results within each school program, while the Southern regions got the worst. Another particularly important variable to be considered is the ESCS (economic, social and cultural status) index, due to its strong relation with student performances demonstrated by several international researches. The ESCS scores are obtained as component scores from a component analysis with zero being the score of an average OECD student and one

Table 1 Mean score in student performance on the reading, mathematics and science scale

	Reading		Mathematics		Science	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
OECD	493	0.5	496	0.5	501	0.5
Italy	486	1.6	483	1.9	489	1.8
Abruzzo	480	4.8	476	6.7	480	5.7
Basilicata	473	4.5	474	4.4	466	3.9
Bolzano	490	3.2	507	3.2	513	2.5
Calabria	448	5.2	442	5.1	443	5.5
Campania	451	6.6	447	7.8	446	6.8
Emilia-Romagna	502	4.0	503	4.7	508	4.8
Friuli Venezia Giulia	513	4.7	510	4.6	524	4.8
Lazio	481	3.9	473	5.5	482	5.0
Liguria	491	9.3	491	9.3	498	9.9
Lombardia	522	5.5	516	5.6	526	5.8
Marche	499	7.3	499	4.5	504	6.5
Molise	471	2.8	467	2.7	469	2.8
Piemonte	496	5.9	493	6.0	501	5.2
Puglia	489	5.0	488	6.9	490	6.3
Sardegna	469	4.3	456	5.2	474	4.5
Sicilia	453	8.3	450	8.8	451	8.2
Toscana	493	4.5	493	5.4	500	5.7
Trento	508	2.7	514	2.5	523	3.6
Umbria	490	5.3	486	4.1	497	5.0
Valle d'Aosta	514	2.2	502	2.3	521	2.6
Veneto	505	5.2	508	5.6	518	5.1
North-West (N-W)	511	3.9	507	4.0	516	4.0
North-East (N-E)	504	2.8	507	2.9	515	2.8
Centre (C)	488	2.6	483	3.2	491	3.0
South (S)	468	3.9	465	4.8	466	4.2
South-Islands (S-I)	456	4.8	451	5.1	454	4.8

the standard deviation across equally weighted OECD countries, where variables taken into account are derived from the student report (highest occupational status of parents, higher parental education expressed in year of schooling and home possessions as a proxy indicator of family wealth).

In the second column of Table 2 the ESCS mean is shown for Italy and the geographical macro-areas. Only the Centre is associated to a positive value of ESCS, meaning that on mean the socio-economic background is higher in this area than in the rest of Italy. This result was also confirmed by the computation of quartiles (not shown here).

Northern regions have negative values but quite close to zero while the Southern regions present largely negative values, meaning that they suffer from an unfavourable background. When looking at the mean score on the reading scale in

Table 2 ESCS mean and mean score on the reading scale based on ESCS national quartiles

	ESCS		Lower quarter		Second quarter		Third quarter		Upper quarter	
	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Italy	−0.12	0.01	441.47	3.01	477.53	2.03	500.51	1.98	526.01	2.14
North-West	−0.06	0.02	464.86	5.56	497.98	5.49	524.84	4.93	548.67	4.50
North-East	−0.03	0.03	457.61	5.33	490.15	3.63	515.31	2.70	546.01	3.84
Centre	0.08	0.03	442.08	4.23	473.57	4.61	499.77	3.25	514.45	3.82
South	−0.32	0.03	436.40	5.82	468.79	4.35	481.26	4.40	507.32	4.79
South-Islands	−0.25	0.04	416.44	8.21	452.88	4.95	469.56	4.86	503.16	6.49

the quarters derived by national quartiles of the ESCS index, it can be noticed that, ESCS ranges being equal, the students in the North, and especially in North-West, outperform students in the Centre and in the South, where Southern-Islands students get the lowest scores.

A multilevel Analysis

The mean score on the reading, mathematics and scientific scales can be considered as a measure of effectiveness both at a country and at a regional level. However, it is well known that much of the variability in student performances can be imputed to school differences (see OECD 2009a, chapter 15) and this aspect deserves a special attention because it is involved with the concept of the system equity. Figure 1 shows the distribution of mean scores on the reading literacy obtained by schools in the five Italian macro-areas. As can be seen, the geographical differences are still present when schools are taken as observations. In fact, schools in the North are associated to a higher mean score with respect to Centre and especially to the South and South-Islands. Furthermore, there is a noticeable variability between schools within the same area (standard deviations range from 69 to 74 PISA score points).

In order to take into account the hierarchical structure of data coming from the PISA survey, we resort to multilevel analysis (Goldstein 1995; Snjiders and Bosker 1999; Bryk and Raudenbush 2002). When data are nested, it is likely that the average correlation between variables measured on observations in the same group will be higher than the average correlation on observations belonging to different groups. For this reason, multilevel models, overcoming the assumption of independence among observations typical of ordinary regression, should be adopted in order to provide reliable and unbiased estimates of linear relationships between variables of interest in the population.

As pointed out in OECD (2009a), the use of linear models without taking into account the way in which students are assigned to schools may provide an incomplete or a misleading representation of efficiency in education systems. On the other hand, multilevel models allow the evaluation of the relative variation in the outcome measures, between students within the same school and between schools (see, e.g., Mantovani and Ricci 2008). Moreover, when considering a third

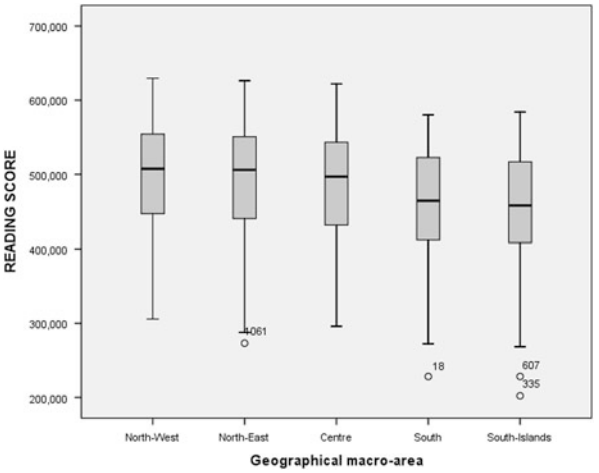

Fig. 1 Boxplots of mean score on the reading literacy obtained by schools in different geographical macro-areas

hierarchical level (e.g. geographical areas or countries) the total variability can be decomposed further.

Firstly, we focus on two-level modelling, considering students as first-level observations and schools as second-level observations. The first step in multilevel analysis is to estimate the so called "empty model", which does not include covariates, and it is equivalent to conduct a mixed analysis of variance. Denoting by Y_{ij} the reading score of the ith student belonging to the jth school, where $i = 1, \ldots, I$ students and $j = 1, \ldots, J$ schools, the empty model can be defined as follows:

$$Y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}, \tag{1}$$

where γ_{00} is the overall intercept, since $\beta_{0j} = \gamma_{00} + u_{0j}$ is the random intercept for school j, u_{0j} is the between school error representing the school departure from the overall intercept and ε_{ij} is the within group error representing the student i departure from the mean score of school j. The variance of the school error (between variance) is $V(u_{0j}) = \tau_{00}$ while the variance of the student error (within variance) is $V(\varepsilon_{ij}) = \sigma^2$. By decomposing the total variance into the group and the residual variances, the intraclass correlation coefficient (ICC) can be computed as

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \tag{2}$$

The ICC represents the expected correlation between two observations of the same group and it is therefore a measure of the degree of homogeneity among observations in the same group. The correlation coefficient expresses the percentage of the total variance that is accounted for by the school; as a consequence, the more the ICC approaches 1, the stronger are the school effect and the need for a multilevel analysis.

Because the multilevel analyses will include the ESCS and the school type as covariates, students with missing ESCS or still in lower secondary school have been deleted. This caused the loss of only 0.68 % of observations, leading to a sample size of 30,695 students. All the multilevel analysis in this section are conducted by using the software HLM (Raudenbush et al. 2008), which is able to handle both sample weights and plausible values in the definition of the response variable.

The results of the estimation of the empty model are presented in Table 3 for Italy, the Italian regions and macro-areas by considering as outcome the reading score. The intercept term is still a measure of efficacy of geographical areas, and again the highest estimates are associated to Northern regions (especially Friuli Venezia Giulia, Lombardia and Trento) while the lowest estimates are recorded for Calabria, Campania and Sicilia. On the other hand, the ICC can be taken as a measure of equity among schools within the same region. The ICC for Italy is about 0.56 meaning that about the 56 % of the variability in the student performance can be explained by school differences. In particular, the less equitable regions are Emilia-Romagna, Lombardia, Marche, Sicilia and Umbria, while the most equitable regions are Abruzzo, Basilicata and Bolzano. Looking at the results in the macro-areas, the North-West seems to be the less equitable area. These results can be deepened by introducing covariates in the model, which are able to reduce both the between and the within variance. Therefore, it will be possible to understand the contribution of the explanatory variables to the decrement in the variability. In particular, we decided to include at the student level the indicator of socio-economic and cultural background (ESCS) and at the school level the school mean ESCS (MU_ESCS) and the school type, recoded into three dummy variables (ACAD for academic, VOC for vocational and VOC_T for vocational training schools) with technical institutes as reference group. The continuous variable ESCS was not centred; in fact, the value "0" is meaningful and corresponds to the mean ESCS in OECD countries.

We consider a model with random intercept and fixed slopes. The first-level model is

$$Y_{ij} = \beta_{0j} + \beta_1 \text{ESCS}_{ij} + \varepsilon_{ij} \tag{3}$$

while the second-level model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{MU\_ESCS} + \gamma_{02}\text{ACAD}_j + \gamma_{03}\text{VOC}_j + \gamma_{04}\text{VOC\_}T_j + u_{0j}. \tag{4}$$

Substituting Eq. (4) into Eq. (3) gives the combined form of the two-level random intercept model with level 1 and level 2 predictors. Analogously to model 1, the between variance is $V(u_{0j}) = \tau_{00}$ while the within variance is $V(\varepsilon_{ij}) = \sigma^2$. Table 4 shows the model estimated parameters.

For this model, the intercept represents the estimated score for a reference student, i.e. with ESCS equal to zero, within a school with ESCS mean equal to zero, and belonging to a technical institute. The reading score for this target student is higher than 500 in all the Northern regions, with the only exception of Liguria, whereas it is particularly low in Calabria, Campania and Sicilia. The net

Table 3 Parameter estimates of the empty model for the reading score

	γ_{00}	τ_{00}	σ^2	ICC
Italy	480.5	5,057.8	4,034.4	0.56
Abruzzo	477.3	3,568.5	3,952.9	0.47
Basilicata	466.8	3,544.9	3,889.5	0.48
Bolzano	490.2	3,790.6	4,381.5	0.46
Calabria	440.0	4,303.6	3,836.7	0.53
Campania	445.9	4,314.8	4,269.5	0.50
Emilia-Romagna	498.8	5,548.3	4,473.9	0.55
Friuli Venezia Giulia	509.3	4,508.4	3,912.2	0.54
Lazio	479.6	4,141.7	4,144.0	0.50
Liguria	485.1	4,534.7	4,358.0	0.51
Lombardia	511.5	4,872.6	3,826.6	0.56
Marche	492.6	4,898.6	3,766.6	0.57
Molise	463.1	3,749.2	3,736.6	0.50
Piemonte	491.6	4,638.9	4,186.6	0.53
Puglia	485.6	3,926.8	3,639.6	0.52
Sardegna	458.9	4,410.4	4,457.3	0.50
Sicilia	450.2	4,743.9	3,928.0	0.55
Toscana	490.1	4,473.2	4,330.8	0.51
Trento	506.3	4,575.1	3,999.4	0.53
Umbria	485.1	5,396.7	4,131.4	0.57
Valle d'Aosta	500.0	4,074.7	3,908.9	0.51
Veneto	502.1	4,119.6	3,804.9	0.52
North-West	503.2	4,890.5	3,977.9	0.55
North-East	501.2	4,704.7	4,095.8	0.53
Centre	484.7	4,456.0	4,140.4	0.52
South	463.0	4,449.5	3,998.1	0.53
South-Islands	450.3	4,571.3	3,981.9	0.53

Note: all intercept estimates are significant at the 1 % level

effect of ESCS (β_1) on the reading performance is largely significant for Italy, Emilia-Romagna and North-East, and it is significant at the 5 % level for the other macro-areas. The effect of the school ESCS mean (γ_{01}) is significant at the 1 % level for Italy, several regions and North-East. Looking at the school type, the impact of being a student in an academic track instead of a technical one (γ_{02}) means an increase of about 34 score points in Italy, and more than 60 points in the Centre and Southern regions. On the other hand, the fact of being a student in a vocational school instead of in a technical one (γ_{03}) means about 43 points less on average, that become about 55 when observing a student in a vocational training track (γ_{04}).

Taking into account socio-economic background and the school type leads to a reduction of the between variance of about 62 % while the reduction of the residual variance is only 4 % for the whole country.

Table 4 Parameter estimates of the two-level random intercept model for the reading score

	γ_{00}	β_1	γ_{01}	γ_{02}	γ_{03}	γ_{04}	τ_{00}	σ^2	ICC
Italy	486.4***	5.1***	44.0***	34.4***	−42.8***	−55.4***	1912.8	4017.0	0.32
Abruzzo	458.3***	2.7	28.2	61.6***	−30.8*	−99.5**	553.2	3947.7	0.12
Basilicata	494.1***	2.7	76.2***	21.7	−40.0**	–	450.9	3884.6	0.10
Bolzano	516.0***	5.2*	17.2	35.2**	−56.5**	−72.9***	822.3	4369.1	0.16
Calabria	425.8***	4.7*	30.9**	69.7***	−24.1	–	858.4	3824.8	0.18
Campania	425.6***	3.4	20.0	81.3***	−25.2	–	1029.3	4261.2	0.19
Emilia-Romagna	500.2***	10.9***	16.2	55.0***	−63.8***	−118.8***	472.2	4401.2	0.10
Friuli Venezia Giulia	519.3***	5.8*	40.6*	36.8**	−67.3***	−97.3***	609.5	3892.8	0.14
Lazio	458.8***	7.0*	−2.7	64.6***	−57.1**	−107.7***	1180.9	4124.3	0.22
Liguria	487.9***	5.4*	36.6	33.6*	−53.2	−51.4*	1841.7	4340.8	0.30
Lombardia	529.3***	5.9*	41.6	19.6	−44.2*	−111.1***	1094.0	3801.8	0.22
Marche	504.8***	2.4	29.4	36.4*	−76.6***	–	1452.0	3768.5	0.28
Molise	473.7***	7.2**	54.9***	21.7	−36.0*	–	1005.1	3694.9	0.21
Piemonte	517.2***	6.3*	35.6	18.2	−78.0***	−54.9***	1161.7	4166.5	0.22
Puglia	498.5***	3.6*	32.6	38.4*	−48.5*	–	1425.6	3630.5	0.28
Sardegna	452.3***	3.9	22.1*	70.4***	−64.4***	–	358.4	4450.8	0.07
Sicilia	435.3***	6.1*	15.3	68.2**	−40.2*	–	1670.8	3899.8	0.30
Toscana	486.0***	5.6	4.1	61.1***	−83.9***	–	784.6	4310.6	0.15
Trento	525.9***	3.6	65.3***	29.4*	−48.4**	−66.6***	488.2	3989.0	0.11
Umbria	471.4***	7.6**	39.4	52.2**	−45.4*	−58.0*	1553.2	4095.7	0.27
Valle d'Aosta	524.3***	2.7	45.7	36.9	−34.8*	−75.4**	917.1	3904.2	0.19
Veneto	524.1***	1.3	54.5**	9.5	−33.1*	−89.0***	1212.0	3803.6	0.24
North-West	522.7***	5.9**	42.5**	17.6	−54.2***	−89.7***	1438.6	3954.8	0.27
North-East	513.9***	5.4***	41.2***	28.7**	−46.7***	−83.4***	1034.9	4077.7	0.20
Centre	473.7***	6.0**	−3.6	61.6***	−70.8***	−118.9***	1310.7	4123.0	0.24
South	452.0***	3.5**	20.3	66.4***	−35.4**	−95.9**	1565.2	3988.3	0.28
South-Islands	438.5***	5.2**	22.0**	66.7***	−40.1***	–	1310.7	3964.7	0.25

Note: *significant at the 10 % level; **significant at the 5 % level; ***significant at the 1 % level

Table 5 Parameter estimates of the three-level empty model for the reading score

	γ_{000}	τ_k	τ_{jk}	σ^2	ICC_k	ICC_{jk}
Regions	481.9	334.78	4,488.6	4,045.1	0.04	0.51

Note: intercept are all significant at the 1 % level

Finally, a last attempt of exploring the different sources of variability in the student performance has been done considering an empty three-level model. In particular, the three-level units have been identified in the regions. The combined three-level model, without covariates follows this formulation:

$$Y_{ijk} = \gamma_{000} + u_{00k} + u_{0jk} + \varepsilon_{ijk}, \tag{5}$$

where $k = 1, \ldots, K$ are the three-level units (regions), Y_{ijk} is the score on the reading scale of the ith student belonging to school jth in the kth region, $V(u_{00k}) = \tau_k$ is the between region variance, $V(u_{0jk}) = \tau_{jk}$ is the variance of schools in the same region and $V(\varepsilon_{ijk}) = \sigma^2$ is the residual (or within) variance. The intraclass correlation can be calculated both at a school and region level by dividing the corresponding variance to the total variance. Results are presented in Table 5.

As can be easily seen, the percentage of variability attributed to regions is only the 4 %, while the 51 % of variability is due to the difference between schools in the same region.

Concluding Remarks

The analyses conducted on data from OECD PISA last edition (2009) showed wide differences in student performances between Italian macro-areas and regions. In particular, students in the North and particularly in the North-West outperformed their colleagues in the Centre and especially in the Southern part of the country. By using a multilevel analysis, it was shown that regions are not different only in effectiveness but also in equity. In fact, much variability in the results can be explained by the between school variance, which is rather high for several regions. Also, by estimating an empty three-level model, it was shown that much of the variability depends on differences between schools within the same region. Analogous results are obtained when analysing mathematical and scientific literacy instead of reading. This result highlights that differences between schools overcome territorial difference. Even if school gaps are widely known by researchers and institutions, the causes should be investigated more in detail, especially in order to suggest ways of improvement.

Acknowledgments We would like to sincerely thank Stefania Mignani and Roberto Ricci for the useful discussions about the data and the results and Aldopaolo Palareti for making possible to map Italian regions by using proficiency levels.

References

Adams, R.J., Wilson, M.R., Wang, W.C.: The mixed coefficients multinomial Logit model. Appl. Psych. Meas. **21**, 1–23 (1997)
Bryk, A.S., Raudenbush, S.W.: Hierarchical Data Analysis. Sage, Newbury Park (2002)
Goldstein, H.: Multilevel Statistical Models. Arnold, London (1995)
Hambleton, R.K., Swaminathan, H.: Item Response Theory: Principles and Applications. Kluwer Nijhoff Publishing, Boston (1985)
INVALSI: La prova nazionale al termine del primo ciclo. Aspetti operativi e prime valutazioni sugli apprendimenti degli studenti (2010a)
INVALSI: Rilevazioni degli apprendimenti SNV. Prime analisi (2010b)
Lord, F.M., Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley, Reading (1968)
Mantovani, D., Ricci, R.: Caratteristiche individuali, caratteristiche delle scuole e competenza in scienze in Emilia-Romagna. In: Gasperoni, G. (ed.) Le Competenze degli Studenti in Emilia-Romagna. I Risultati di PISA 2006, pp. 197–226. Il Mulino, Bologna (2008)
OECD: PISA Data Analysis Manual. OECD, Paris (2009a)
OECD: PISA 2009 Assessment Framework. Key Competencies in Reading, Mathematics and Science. OECD, Paris (2009b)
OECD: PISA 2009 Results: What Students Know and Can Do. Student Performance in Reading, Mathematics and Science (Volume I). OECD, Paris (2010)
OECD: PISA in focus. Improving performance: leading from the botton. OECD, Paris (2011)
Raudenbush, S.W., Bryk, A.S., Cheong, Y.F.: HLM 6.06 for Windows [computer software]. Scientific Software International, Lincolnwood (2008)
Snjiders, T., Bosker, R.: Multilevel Analysis. Sage, London (1999)

Some Experimental Results on the Role of Speed and Accuracy of Reading in Psychometric Tests

Isabella Morlini, Giacomo Stella, and Maristella Scorza

Abstract

According to the Italian Parliament act (n. 170/2010) that recognizes dyslexia as a physical disturbance, of neurobiological origin, dyslexic children in primary school should be early recognized, in order to assess a targeted intervention within the school and to start a teaching that respects the difficulties in learning to read, to write, and to perform calculations. Screening procedures inside the primary schools aimed at detecting children with difficulties in reading are not so common in Italy as in other European countries. Nevertheless, screening procedures are of fundamental importance for guaranteeing an early detection of dyslexic children and reducing both the primary negative effects—on learning—and the secondary negative effects—on the development of the personality—of this disturbance. In this study we analyze the validity, from a statistical point of view, of a screening procedure recently proposed in the psychometric literature (Stella et al., SPILLO Strumento per l'identificazione della lentezza nella lettura orale. Software per la verifica delle abilit di lettura dei bambini della scuola primaria (dalla prima alla quinta). Giunti Scuola, Florence, 2011). This procedure is very fast (it is exactly one minute long), simple, cheap and can be dispensed by teachers without psychometric experience. On the contrary, the currently used tests are much longer and must be provided by skilled teachers. These two major flaw prevent the widespread use of these tests. If the new procedure is found to

I. Morlini (✉)
Department of Economics, University of Modena & Reggio Emilia, Via Berengario 51, 41.121 Modena, Italy
e-mail: isabella.morlini@unimore.it

G. Stella • M. Scorza
Department of Social Sciences, University of Modena & Reggio Emilia, Viale Allegri, 42.121 Reggio Emilia, Italy
e-mail: giacomo.stella@unimore.it; maristella.scorza@unimore.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_27,

be reliable, it can be provided to each student in primary school and it can also be repeated in time, in order to monitor the children difficulties. The validity of the procedure and the benchmark with two currently used tests are studied on the basis of the results of a survey on about 1,500 students attending primary school.

Introduction

The act of Parliament n. 170 (approved the 8th October 2010), on "the new statutory law for learning disorders affecting the scholastic population" states that dyslexia is a physical disturbance, of neurobiological origin, which makes it very difficult to learn to read, to write, and to perform calculations for intelligent children who do not have any other types of disorder. According to this act, teaching a dyslexic child in the school should respect his pace and learning methods and should include a system of assessment that takes into account his different performances. The early detection of dyslexic children in primary school becomes then of fundamental importance. Procedures for recognizing dyslexic children are based on reading performance. Indeed, even though Italians use a shallow orthography which facilitates reading, Paulesu et al. (2001) found that an Italian dyslexic reads better than a French and an English dyslexic, but he performs significantly worse than a non-dyslexic Italian reader. The aim of this paper is to study the validity, from a statistical perspective, of a screening procedure recently proposed in the psychometric literature (Stella et al. 2011). This procedure is very fast (it is exactly one minute long), simple and can be dispensed by teachers without psychometric experience. The reading tests currently used in Italy, in primary school, are much longer and must be dispensed by skilled teachers. Drawing from the results obtained by administering the screening procedure and, as a benchmark, two currently used reading tests, to about 1,500 students in primary school, in Italy, the purpose of this work is twofold:

– To study the empirical distribution of the variables measuring speed and accuracy of reading in the screening procedure and in the benchmark tests. This study allows us to discuss the validity of the threshold values used for classifying the student's performance as impaired or as not impaired.
– To study the validity of the screening procedure, through an explorative factor analysis and through the estimate of the internal consistency.

The paper is organized as follows. In Sect. 2 we briefly illustrate the screening procedure and the tests used as benchmarks. In Sects. 3 and 4 we report univariate and bivariate analyses of the variables measuring the reading performances, we analyze the empirical distribution of these variables and we discuss the choice of the threshold values currently used for classifying the student's performance. In Sect. 5 we analyze the validity of the screening procedure. Finally, in Sect. 6 we give some concluding remarks.

The Screening Procedure and the Benchmark Tests

The new screening procedure for students attending primary school, in Italy, is called SPILLO. It is implemented within a software and thus the results of the screening are immediately available after the implementation. The student is asked to read a text for exactly 1 min. The examiner clicks the word spelled wrongly on the screen and, after 1 min, the software emits a sound that indicates the end of the procedure. The examiner clicks the last word read by the student and the software computes the number of words, the number of syllables read in a second, the number of errors and, eventually, the z-scores. Since the evaluation of reading ability should take into account both the single word reading process (which is an explicit and automatic process) and the lexical anticipation (which is a higher level process), in SPILLO the reading performances are assessed in a text rather than in a list of words or nonwords (words without a meaning). The text chosen is a story composed of 181 words. The variables measuring the reading performance are:

Y_1: number of words read in a minute,
Y_2: number of syllables per second read (in a minute).
Y_3: number of wrong spelling in a minute.

The benchmark tests are two currently used "paper and pencil" tests. While the student reads, the examiner times the reading and makes a note of the mistakes. Then, the examiner classifies the student as an impaired reader or as a not impaired reader, on the basis of the normative threshold values. This procedure is very simple for experts but it may result rather difficult for examiners who lack a statistical background. Indeed, results are usually affected by many errors due to erroneous calculus or wrong interpretation of the results. In the first test the student is asked to read a list of words and in the second test a list of nonwords. These two lists have been introduced and studied by Sartori et al. (1995, 2007). In the two benchmark tests the variables measuring the reading performance are:

X_1: time (in seconds) in reading the list of words,
X_2: number of syllables per second read in the list of words,
X_3: number of wrong spellings in reading the list of words,
X_4: time (in seconds) in reading the list of nonwords,
X_5: number of syllables per second read in the list of nonwords,
X_6: number of wrong spellings in reading the list of nonwords.

The new screening procedure and the two benchmark tests have been provided to 1,469 students in elementary school, in the Lombardia and the Emilia Romagna regions (Northern Italy). The tests have been administered to students attending classes II–V in February and to students attending class I in May. Since Italian is a language with transparent or shallow orthography, where the letters of the alphabet, alone or in combination, are in most instances uniquely mapped to each of the speech sounds occurring in the language, at the end of the first school year students are in general able to read. The sample does not include foreign students.

The composition of the sample is: 333 students attending class I, 384 students attending class II, 200 students attending class III, 276 students attending class IV, and 276 students attending class V.

Reading Speed

In Table 1 we list the values of some univariate statistics of the variables measuring the reading speed (X_1, X_2, X_4, X_5, Y_1, Y_2). Figure 1 shows the empirical distributions of these variables through boxplots. Performances in reading speed improve from class I to class V: both the median and the mean values of X_1 and X_3 decrease while the mean and the median values of X_2, X_4, Y_1, and Y_2 increase. Variables measuring the number of words and the number of syllables read in a second have a similar pattern: the average values of Y_1 and Y_2 across the five classes have a behavior similar to the average values of X_2 and X_5. Dispersion, measured by the coefficient of variation, always decreases with the grade level. In the time of reading (X_1 and X_4) and in Y_1 and Y_2, a drop of the coefficient of variation corresponding to class III is evident. This drop in not present in X_2 and X_5. The larger variability in classes I, II, and III may be explained by considering that many covariates (like the cultural level, the experiences in day nursery, etc...) have a great influence on the reading performances. From class III, in general, these covariates become irrelevant and the scholastic population becomes more homogeneous.

Variables measuring the speed of reading in the benchmark tests (X_1, X_2, X_4, X_5) have a positive skew and present outlying values higher than $x_{0.75} - 1.5(x_{0.75} - x_{0.25})$, in all classes. These characteristics are desirable for X_1 and X_4 that have a "positive direction of pathology" (impaired readers are children with high values in these variables), but not for X_2 and X_5 that have a "negative direction of pathology" (impaired readers are children having small values in these variables). Y_1 and Y_2 have a positive skew in classes I and II but a negative skew in classes II, IV, and V. In IV and V these variables have outlying values smaller than $x_{0.25} + 1.5(x_{0.75} - x_{0.25})$.

The currently used threshold for X_1, X_2, X_4, X_5 are based on the assumption of normality. The thresholds are used for classifying students as normal readers or impaired readers. They have been specified on the basis of the mean and the variance, assuming a normal distribution (Sartori et al. 1995, 2007). The thresholds have been obtained as $\mu + 2\sigma$ (for X_1 and X_4) and as $\mu - 2\sigma$ (for X_2 and X_5), where μ indicates the mean and σ the standard deviation, considering that in a Gaussian distribution these values exclude about 2 % of the population. The estimated values of μ and σ, reported in Sartori et al. (2007) and currently used as normative values in the tests, have been estimated for the classes II, III, IV, and V on a very small sample. Using the T-test for the means and the non-parametric test of Levene for the variances, they result significantly different ($\alpha = 0.05$) from the means and the variances obtained in our study and reported in Table 1. Moreover, with the three non-parametric tests of Shapiro–Wilk, Anderson–Darling, and Jarque–Bera, the null hypothesis of Gaussian distribution of the variables X_1, X_2, X_4, X_5, Y_1, Y_2, in each

Table 1 Univariate statistics obtained in the sample for variables measuring the reading speed. Values in bold are the selected thresholds for the screening procedure

	X_1					X_4					Y_1				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
N.	333	384	200	276	276	333	384	200	276	276	333	384	200	276	276
Min	92	69	52	50	43	58	49	33	34	31	8	21	48	9	57
Max	1148	595	289	348	206	542	311	210	234	118	127	157	164	180	180
$x_{0.05}$	145	96	70	59	54	89	68	52	49	38	16	34	64	78	96
$x_{0.25}$	232	135	90	73	64	120	89	69	59	50	30.0	54.0	77.0	108.0	127.0
$x_{0.50}$	314	177	110	89	74	151	110	84	72	59	41.0	66.0	97.0	125.0	146.0
$x_{0.75}$	424	234	146	108	90	204	137	98	86	73	53.0	81.0	115.3	140.0	164.0
$x_{0.95}$	723	381	209	153	121	342	206	157	109	97	**72**	**113**	**138**	**170**	**180**
Mean ($\bar{x}$)	360	198	121	96	79	173	120	88	74	63	43.0	69.0	97.7	123.5	143.7
S	190.5	88.9	43.7	34.2	23.5	82.4	45.4	29.8	23.8	17.8	19.9	23.9	24.3	26.9	25.3
$S/\bar{x}$	0.53	0.45	0.36	0.36	0.30	0.48	0.38	0.34	0.32	0.28	0.46	0.35	0.25	0.22	0.18
Skewness	1.66	1.49	1.30	2.53	1.84	1.76	1.64	1.45	2.23	0.61	1.21	0.64	−0.23	−0.42	−0.76
	X_2					X_5					Y_2				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Min	0.24	0.47	0.97	0.80	1.36	0.23	0.41	0.60	0.54	1.07	0.30	0.78	1.77	0.32	2.07
Max	3.04	4.06	5.38	5.60	6.51	2.17	2.57	3.82	3.71	4.06	4.23	5.25	5.53	5.98	5.98
$x_{0.05}$	0.39	0.73	1.34	1.83	2.31	0.37	0.61	0.80	1.16	1.30	**0.58**	**1.27**	**2.30**	**2.80**	**3.32**
$x_{0.25}$	0.66	1.20	1.92	2.59	3.11	0.62	0.92	1.29	1.47	1.73	1.10	1.97	2.77	3.63	4.23
$x_{0.50}$	0.89	1.58	2.56	3.16	3.78	0.83	1.15	1.51	1.76	2.14	1.48	2.38	3.33	4.18	4.87
$x_{0.75}$	1.21	2.08	3.12	3.84	4.38	1.05	1.42	1.83	2.14	2.52	1.95	2.90	3.86	4.65	5.53
$x_{0.95}$	1.93	2.92	4.00	4.75	5.19	1.42	1.85	2.43	2.82	3.32	2.65	3.82	4.55	5.70	5.98
Mean ($\bar{x}$)	0.98	1.67	2.59	3.20	3.77	0.86	1.18	1.59	1.85	2.17	1.56	2.46	3.36	4.16	4.82
S	0.48	0.66	0.81	0.89	0.92	0.33	0.37	0.49	0.53	0.62	0.68	0.77	0.73	0.81	0.79
$S/\bar{x}$	0.49	0.40	0.31	0.28	0.24	0.38	0.32	0.31	0.29	0.29	0.44	0.31	0.22	0.19	0.16
Skewness	1.48	0.74	0.34	0.16	0.03	0.54	0.45	0.83	0.72	0.69	1.06	0.45	−0.09	−0.34	−0.69

class, is rejected, even for $\alpha = 0.001$. The normative thresholds, in our sample, lead to the following percentages of students classified as impaired readers:

	X_1 (%)	X_2 (%)	X_4 (%)	X_5 (%)
Class II	7.8	0.5	5.7	0.0
Class III	3.0	0.0	4.5	0.5
Class IV	2.5	1.1	1.4	0.7
Class V	2.2	0.7	1.1	0.0

These percentages vary greatly not only across the classes but also across the variables. In class II, for example, 30 students are classified as impaired readers with X_1 and only 2 with X_2. Quite all the percentages are far from the expected value, equal to the estimated percentage of dyslexic students in the Italian population (about 4 %). Due to the non-normality of the variables and the presence of outliers, thresholds for Y_1 and Y_2 have been set equal to the percentile $x_{0.5}$ obtained in our

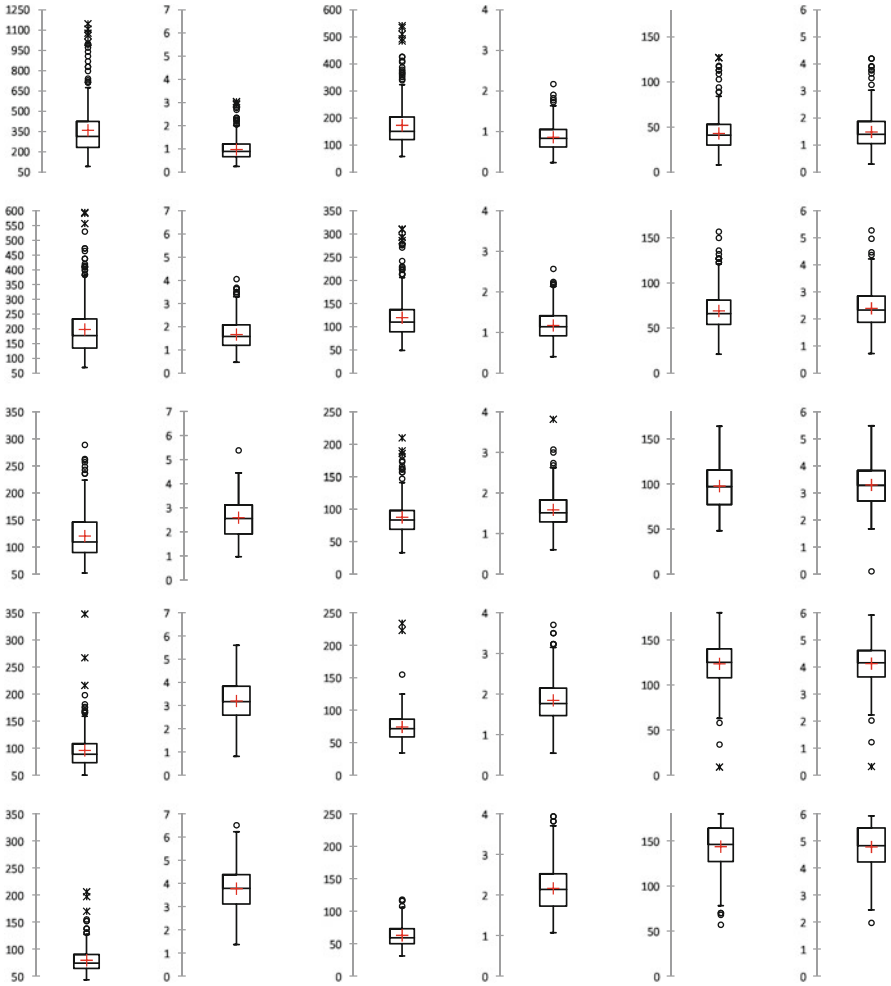

Fig. 1 Boxplots of the following variables (from the *left*): X_1,X_2, X_4, X_5, Y_1, Y_2. The *first line* reports boxplots in class I, the *second line* in class II, the *third line* in class III, the *fourth line* in class IV, and the *fifth line* in class V

sample (the values are reported in bold in Table 1) for discriminating a percentage of people slightly higher than the expected percentage. As a matter of fact, the procedure is not intended as a diagnostic test for learning disorders but as a screening for detecting students with heavy difficulties in reading. The causes of these difficulties are to be defined by subsequent more detailed analyses. Future studies are necessary to explore the percentages of students classified as impaired readers with these thresholds in a new sample of students.

To investigate the validity of the screening procedure, we may analyze the pairwise correlations between the variables Y_1 and Y_3 and the variables used in

Table 2 Correlation matrices for variables measuring reading speed in classes I and IV

	Class I							Class IV						
	X_1	X_2	X_4	X_5	Y_1	Y_2		X_1	X_2	X_4	X_5	Y_1	Y_2
X_1	1	−0.79	0.91	−0.81	−0.72	−0.73	X_1	1	−0.89	0.84	−0.71	−0.76	−0.75
X_2	−0.79	1	−0.72	0.9	0.91	0.91	X_2	−0.89	1	−0.77	0.83	0.77	0.76
X_4	0.91	−0.72	1	−0.86	−0.66	−0.66	X_4	0.84	−0.77	1	−0.88	−0.61	−0.61
X_5	−0.81	0.9	−0.86	1	0.82	0.82	X_5	−0.71	0.83	−0.88	1	0.61	0.6
Y_1	−0.72	0.91	−0.66	0.82	1	1	Y_1	−0.76	0.77	−0.61	0.61	1	1
Y_2	−0.73	0.91	−0.66	0.82	1	1	Y_2	−0.75	0.76	−0.61	0.6	1	1

the benchmark tests. The correlation matrices for class I and IV are reported in Table 2. The matrices obtained in class II, III, and V look very similar. The pairwise correlations are all significantly different from zero ($\alpha = 0.01$). Even though the transformation from X_1 to X_2 and from X_3 to X_4 is not a linear one, the values of these pairs of variables are highly correlated. The transformation from Y_1 to Y_2 is not a perfect linear transformation (since the words have different numbers of syllables) but the correlation is nevertheless equal to 1. The fact that Y_1 to Y_2 are highly correlated with X_1, X_2, X_4, and X_5 gives evidence that all these variables are a measure of the same aspect of the phenomenon dyslexia.

Reading Accuracy

In Table 3 we list the values of some univariate statistics in the variables measuring the accuracy of reading (X_3, X_6, Y_3). Figure 2 shows the empirical distributions of these variables through boxplots. As well as the variables measuring the speed of reading, X_3, X_6, and Y_3 have an empirical distribution which is asymmetric and far from the Gaussian. While the mean and the median values of X_3 and X_6 have a decreasing pattern, from class I to class V, the mean and the median values of Y_3 are roughly constant across classes. This different pattern is due to the fact that the time in the screening procedure is always equal to 1 min, while it depends on the ability of the student in the benchmark tests. In the screening, if one student increases the performance from one class to the subsequent class, he increases the reading speed without penalizing the reading accuracy. Outliers are all in the "direction of pathology" and this is a desirable property. The normative threshold values for X_3 and X_6 are the 95th percentiles obtained in the study of Sartori et al. (2007). These values are similar to $x_{0.95}$ obtained in our sample (reported in Table 3). The percentages of students classified as impaired readers with the currently used thresholds are as follows: class II: 2.9 % (X_3), 4.4 % (X_6), class III: 1.5 % (X_3), 4.0 % (X_6), class IV: 4.3 % (X_3), 9.4 % (X_2), class V: 3.6 % (X_1), 4.0 % (X_2). These percentages are in general in agreement with the expected value. In the screening, the threshold valued in Y_3 is the value $x_{0.95}$ obtained in our sample. While X_3 and X_6 are highly correlated, Y_3 is only slightly positive correlated with X_3 or

Table 3 Univariate statistics obtained in the sample for variables measuring the reading accuracy. Bold values are the selected thresholds for the screening procedure

	X_3					X_6					Y_3				
	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Min	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Max	84	55	18	19	22	38	31	20	24	27	9	10	6	11	11
$x_{0.25}$	5.0	4.0	2.0	1.0	1.0	4.0	5.0	3.0	3.0	2.0	0.00	0.50	0.50	0.88	0.50
$x_{0.50}$	10.0	6.0	4.0	3.0	2.0	8.0	7.0	6.0	5.0	4.0	1.00	1.50	1.50	1.00	1.00
$x_{0.75}$	17.0	10.3	6.0	5.0	4.0	13.0	10.0	8.0	8.0	6.0	2.00	2.50	2.00	2.50	2.00
$x_{0.95}$	38	19	11	10	9	24	17	14	14	12	**4.0**	**5.0**	**4.0**	**4.0**	**4.0**
Mean ($\bar{x}$)	13.57	7.96	4.48	3.61	2.8	9.83	8.16	6.29	5.7	4.8	1.30	1.76	1.41	1.69	1.48
S	12.34	6.78	3.56	3.34	3.1	7.20	5.13	4.02	4.4	4.1	1.42	1.51	1.18	1.52	1.53
$S/\bar{x}$	0.91	0.85	0.79	0.93	1.12	0.73	0.63	0.64	0.76	0.84	1.09	0.86	0.84	0.90	1.03
Skewness	2.07	2.62	1.29	1.63	2.55	1.29	1.18	1.06	1.32	1.92	1.80	1.35	0.97	1.81	2.29

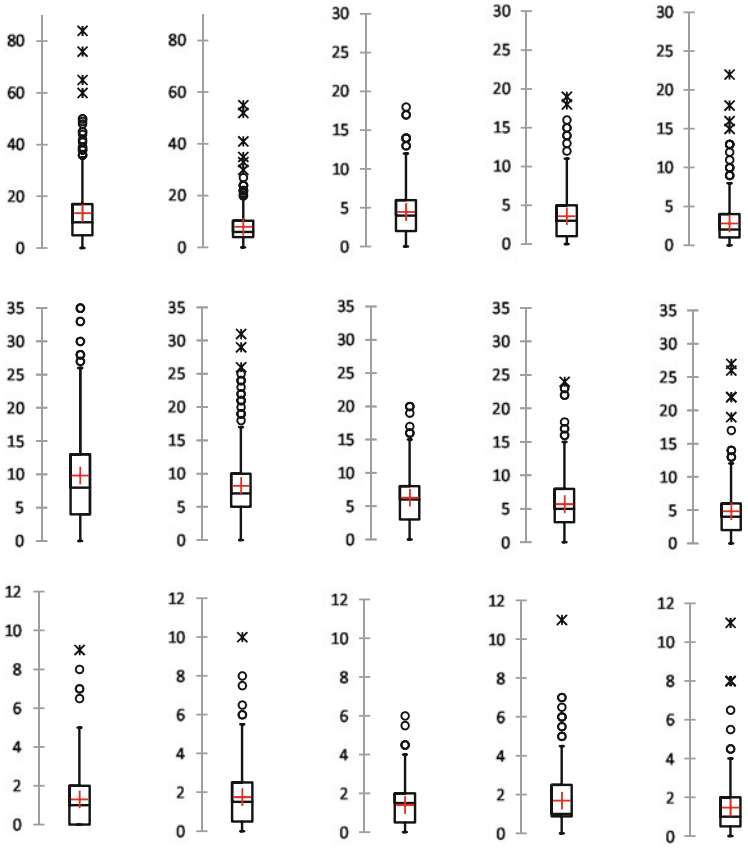

Fig. 2 Boxplots of X_3 (*first line*), X_6 (*second line*), and Y_3 (*third line*). Class I is in the first column, class II in the second, class III in the third, class IV in the fourth, and class V in the fifth column

Table 4 Correlation matrices for variables measuring the reading accuracy

	Class I			Class II			Class III			Class IV			Class V		
	X_3	X_6	Y_3	X_3	X_6	Y_3	X_3	X_6	Y_3	X_3	X_6	Y_3	X_3	X_6	Y_3
X_3	1	0.8	0.37	1	0.69	0.43	1	0.68	0.31	1	0.61	0.34	1	0.61	0.35
X_6	0.8	1	0.38	0.69	1	0.39	0.68	1	0.33	0.61	1	0.28	0.61	1	0.3
Y_3	0.37	0.38	1	0.43	0.39	1	0.31	0.33	1	0.34	0.28	1	0.35	0.3	1

with X_6 (Table 4). This is due to the fact that the time is fixed in the screening. As explained previously, the number of mistakes has a different pattern from the number of mistakes in a reading test where the time depends on the ability of the subject.

Internal Consistency

An explorative factor analysis has been performed in order to investigate the multivariate relationships among variables used in the benchmark tests and in the screening procedure. The analysis, performed on the correlation matrix (reached with the values of all variables in all classes) shows the presence of two main latent orthogonal factors both with the principal component (PC) and with the common factors (CF) method. The first factor is highly correlated with variables measuring speed and, to a less degree, with X_3 and X_6. With the PC method, the eigenvalue of this factor is equal to 6 and the percentage of explained variance is 66.7 %. With the FC method, the eigenvalue is 5.83 and the percentage of variance 64.8 %. The second factor is highly correlated with Y_3 and, to a less degree, with the other variables measuring accuracy (X_3 and X_6). With the PC method, the eigenvalue of this factor is 1.28 and the percentage of explained variance is 14.3 %. With the CF method, the eigenvalue is 0.86 and the percentage of variance 9.6 %. Figure 3 and Table 5 summarize results obtained with the PC method. Drawing from these results, with the belief that speed and accuracy of reading are two different aspects of dyslexia, we estimate the degree to which the set of variables X_1, X_2, X_4, X_5, Y_1, Y_2 measures a single unidimensional latent construct (the speed of reading) and the set of variables (X_3, X_6, Y_3) measures an other unidimensional latent construct (the accuracy of reading). We estimate the internal consistency of each set of variables by means of the coefficient ω (McDonald 1999; Zinbarg et al. 2005), considering the correlation matrix. For the variables regarding speed, $\omega = 0.86$. For the variables regarding accuracy, $\omega = 0.64$. Since these variables have all positive pairwise correlations we may also calculate the α coefficient of Cronbach (1951) and the ρ^* coefficient of Brown (1910) used in the psychometric literature. We obtain $\alpha = 0.85$ and $\rho^* = 0.70$. Regarding the speed of reading, if we select variables having positive pairwise correlations (namely, X_2, X_5, Y_1, Y_2), ω is 0.94 and both the α coefficient and the ρ^* coefficient are 0.98. Even though, in each set, the standardized variables are not τ-equivalent and α gives an overestimation

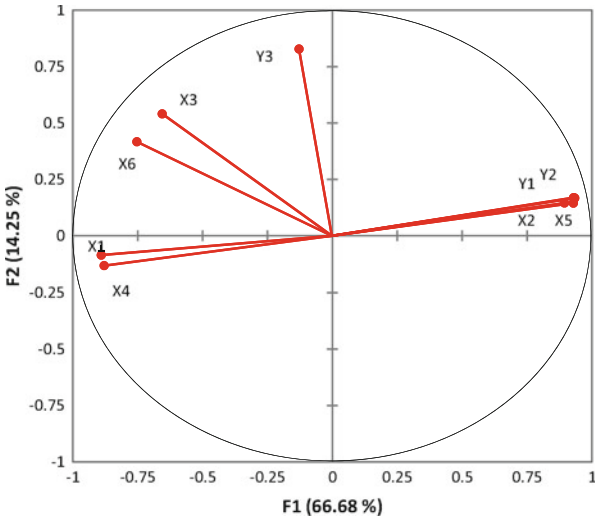

Fig. 3 Biplot resulting from factor analysis applied to the correlation matrix. The factors are extracted with the principal component method and are unrotated

Table 5 Correlations between factors and variables

Factor	X_1	X_2	X_3	X_4	X_5	X_6	Y_1	Y_2	Y_3
F_1	−0.89	0.93	−0.65	−0.88	0.90	−0.75	0.93	0.94	−0.13
F_2	−0.09	0.15	0.54	−0.13	0.15	0.42	0.17	0.17	0.83

The factors are extracted with the principal component method and are unrotated

of the internal consistency, all indexes show high inter-correlation among variables belonging to each set.

Since the screening procedure measures two latent constructs of the dyslexia phenomenon, to evaluate its reliability, we may consider the percentage of variance explained by the first two factors. With both the PC and the PF method of extraction, this percentage is 99.97 % and indicates a very high reliability.

Conclusions

In this paper we have studied the validity, from a statistical perspective, of a new screening procedure proposed in the psychometric literature for the early detection of dyslexia in primary school. On the basis of the results obtained in a sample of 1,469 students we have shown that this screening procedure is able to measure speed and accuracy of reading as well as the currently used tests. The analysis of the empirical distribution of the variables measuring the reading performance in the tests has shown that the normative thresholds, used for classifying a student as a normal reader or as an impaired reader, do not seem to be trustworthy. Indeed, in

our study the variables are found to be far from the normal distribution whereas the assumption of normality has been used to defined these thresholds. Moreover, the means and the variances that we have obtained in our sample are statistically different from the means and the variances used as normative values.

References

Brown, W.: Some experimental results in the correlation of mental abilities. Br. J. Psychol. **3**, 296–322 (1910)

Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**(3), 297–334 (1951)

McDonald, R.P.: Test Theory: A Unified Treatment, pp. 90–103. Erlbaum, Mahwah (1999)

Paulesu, E., Demonet, J.F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S.F., Cossu, G., Habib, M., Frith, C.D., Frith, U.: Dyslexia: cultural diversity and biological unity. Sciences **291**, 2165–2167 (2001)

Sartori, G., Job, R., Tressoldi, P.E.: DDE Batteria per la valutazione della dislessia e della disortografia evolutiva. Giunti O.S., Firenze (1995)

Sartori, G., Job, R., Tressoldi, P.E.: DDE-2 Batteria per la valutazione della dislessia e della disortografia evolutiva - 2. Giunti O.S., Firenze (2007)

Stella, G., Scorza, M., Morlini, I.: SPILLO Strumento per l'identificazione della lentezza nella lettura orale. Software per la verifica delle abilit di lettura dei bambini della scuola primaria (dalla prima alla quinta). Giunti Scuola, Florence (2011)

Zinbarg, R., Revelle, W., Yovel, I., Li, W.: Cronbach's, Revelle's, and McDonald's: their relations with each other and two alternative conceptualizations of reliability. Psychometrika **70**, 123–133 (2005)

A Propensity Score Matching Method to Study the Achievement of Students in Upper Secondary Schools

Giulia Roli and Luisa Stracqualursi

Abstract

In the paper, we investigate the effects of family characteristics on the achievement of students in the first year of the upper secondary schools of the province of Bologna. In particular, we focus our attention on the number of siblings as potential causal factor influencing the outcome. We employ a matching strategy based on propensity score to create treatment groups, corresponding to the values of the factor under study, with the same distribution of observed covariates. As a result, students are stratified in blocks according to the propensity score to obtain estimates of the average treatment effect using nearest neighbour matching. In order to further compare the achievements of students of upper secondary schools in the city of Bologna with those in the other towns of the province, we show that valid inference is assured by controlling for family characteristics whose influence on the outcome has been previously assessed.

Introduction

The investigation of the factors which may influence the achievement of students in the different levels of their education is a crucial topic in observational studies on individual learning experiences and, more generally, in evaluation researches of educational field. The substantive goals are to discern groups of students characterized by specific features which are likely to affect the educational attainments and

G. Roli (✉) • L. Stracqualursi
Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy
e-mail: g.roli@unibo.it; luisa.stracqualursi@unibo.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_28,

to help teachers to properly counsel students during the school year and towards the subsequent educational levels. Despite in the last decades the free access to the learning services is assured, several studies assert that family background, in terms of socio-economic indicators, and family structure affect the educational attainment and the educational achievement of students with different impacts among countries and levels of education (Erikson et al. 2005; Lauer 2003). Therefore, collecting more accurate information on family, such as educational and job qualification of parents, income and marital status, as well as students features and achievements, ensures a valid framework to investigate different aspects of the phenomenon and, as a final result, to properly define local educational policies.

In this paper, we aim at evaluating the effects of family characteristics on the achievement of the students, in terms of success or failure at the end of the school year, referring to data on students in the first year of the upper secondary schools of the province of Bologna in year 2007/2008. We consider the number of siblings as potential causal factor influencing the outcome. Indeed, in some previous analyses Mignani et al. (2011a,b) this has been emerged as a covariate highly associated with the educational attainment of the same group of students. In order to further compare the achievements of students of upper secondary schools in the city of Bologna with those in the other towns of the province, we show that valid inference is assured by controlling for family characteristics. In the analysis, we employ a matching strategy based on propensity score (Rosenbaum and Rubin 1983) to create treatment groups, corresponding to the values of the factor under study, with the same distribution of observed covariates.

Data

In order to increase the information which are commonly available, the province of Bologna, through the *Osservatorio della Scolarità* agency, experimented a more detailed collection of data during the enrollment at the school year 2007/2008. Indeed, family socio-cultural and economic conditions, in addition to the educational and demographic characteristics of the students, have been gathered through a specific questionnaire.

In this paper, we consider a sample of 2,012 students registered at the first year of the upper secondary schools of the province of Bologna. Some descriptive statistics on available information are summarized in Table 1.

We define the outcome of interest Y as the success ($Y = 1$) or the failure ($Y = 0$) at the end of the first school year. In particular, at the end of the year, 1,689 students (83.95 %) passed at the second year and 323 (16.05 %) failed. As potential causal factor influencing the outcome, we denote by T_1 the number of siblings. In particular, students with one sibling at most are grouped and denoted by $T_1 = 0$ and students with two siblings and over are indexed by $T_1 = 1$. We further aim

Table 1 Descriptive statistics

Variable		Number of students	%
Sex	Male	986	49.01
	Female	1,026	50.99
Nationality	Italian	1,890	93.94
	Other	122	6.06
Type of school	Academic	1,166	57.95
	Vocation	339	16.85
	Technical institute	507	25.2
Type of employment of the head of the household	Manager	222	11.03
	Teacher/serviceman/office worker/skilled worker	828	41.15
	Unskilled worker/domestic worker	239	11.88
	Enterpreneur/freelancer	283	14.07
	Craftsman/merchant/farmer	253	12.57
	Unemployed/homemaker	41	2.04
	Pensioner/disabled/deceased	53	2.63
	Missing	93	4.62
Educational qualification of the mother	Primary school	71	3.53
	High school	430	21.37
	Upper secondary school	1,011	50.25
	Degree	410	20.38
	Missing	90	4.47
Educational qualification of the father	Primary school	94	4.67
	High school	533	26.49
	Upper secondary school	869	43.19
	Degree	398	19.78
	Missing	118	5.86
Marital status of parents	Married/cohabitant	1,444	71.77
	Separated/divorced/widower	278	13.82
	Missing	290	14.41
Number of siblings	0	527	26.19
	1	970	48.21
	2	223	11.08
	3	52	2.58
	4	4	0.2
	6	1	0.05
	Missing	235	11.68

at comparing the achievements of students of upper secondary schools in the city of Bologna with those in the other towns of the province, controlling for the family structure and background. We thus define an additional treatment T_2, which assumes values 1 if the students are enrolled into a school of the city, and 0 otherwise.

Methods

Under a causal inference framework, the number of siblings (T_1) and the school location (T_2) are separately regarded as treatments whose effect would be estimated. In observational studies, we have no control over the assignment of the treatment to units. Therefore, we focus on a setting to select a sample where the treatment and control samples are more balanced, i.e. the marginal covariate distributions in the two treatment arms are similar. In particular, we use propensity score to select the observed pre-treatment covariates mostly correlated with the treatments (Rosenbaum and Rubin 1983) and then match treated units with controls to obtain estimates of the average treatment effect on the treated (ATT) (Rubin 1974) using nearest neighbour method. In other words, the propensity score, defined as the conditional probability of receiving the treatment T given a set of covariates $\mathbf{X}$, $e(\mathbf{x}) = Pr(T = 1|\mathbf{X} = \mathbf{x})$, is here used as a balancing score, $b(\mathbf{x})$, i.e. a function of the covariates such that $\mathbf{X}_i \perp T_i|b(\mathbf{X}_i)$ for each unit i in the sample.

In the specification of the propensity score for each treatment, the following covariates available from the data collection are considered: sex, nationality and type of school (academic, vocation or technical institute) of pupils, type of employment of the head of the household, educational qualification of both parents, marital status of parents, number of siblings (only for T_2).

In order to assess the covariate balance between the treated and control samples after the matching, we adopt several measures based on the estimated propensity score $\hat{e}(x)$ (Imbens and Rubin 2014). Indeed, any imbalance in the covariate distribution leads to a difference in the distribution of the propensity scores by treatment arm. A first measure is represented by the normalized difference (Imbens and Wooldridge 2009), which can be estimated as

$$\hat{\Delta}_{01} = \frac{\overline{e}_1 - \overline{e}_0}{\sqrt{S_1^2 + S_0^2}} \tag{1}$$

where $\overline{e}_1$ and $\overline{e}_0$ are the average of the estimated propensity scores for the treatment and control group, respectively, and S_1^2 and S_0^2 denote the conditional within-group sample variances of the estimated propensity score values. In addition to comparing the differences in location in the two distributions, one may wish to compare measures of dispersion in the two distributions through the logarithm of the ratio of standard deviations, which can be estimated as

$$\ln(S_1) - \ln(S_0). \tag{2}$$

Moreover, we can investigate how what fraction of the treated (control) units have propensity score values that are in the centre of the distribution of the propensity score values for the controls (treated), for instance by calculating the probability mass of the propensity score distribution for the treated that is within, say, the 0.975

and 0.025 quantiles of the propensity score distribution for the controls. An estimate of this measure of overlap is

$$\hat{\pi}_1 = \hat{F}_1\left(\hat{F}_0^{-1}(0.975)\right) - \hat{F}_1\left(\hat{F}_0^{-1}(0.025)\right) \tag{3}$$

for treated units and

$$\hat{\pi}_0 = \hat{F}_0\left(\hat{F}_1^{-1}(0.975)\right) - \hat{F}_0\left(\hat{F}_1^{-1}(0.025)\right) \tag{4}$$

for controls, where $\hat{F}_1(\cdot)$ and $\hat{F}_0(\cdot)$ are the empirical distribution function of X in the subpopulation of treated and control units, respectively, and $\hat{F}_1^{-1}(q)$ and $\hat{F}_1^{-1}(q)$ ($q = 0.975, 0.025$) are their inverse.

Finally, as a graphical way to assess balancing, it may be useful to generate histograms of the estimated propensity scores for the treated and the controls with bins corresponding to the strata constructed for the estimation of propensity scores. Ideally, we would like to reach an equal frequency of treated and control in each bin.

Results

The main results are reported in Table 2. The propensity score matching and the ATT estimation have been obtained using the Stata program by Becker and Ichino (2002).

As far as the number of siblings T_1 is concerned, they show the matching strategy creates 943 controls (i.e. students with one sibling at most) to be compared with 280 treated pupils (with more than one sibling). The balance in the covariate distributions is excellent. Indeed, for the propensity score the normalized difference in covariate means $\hat{\Delta}_{01}$ is 0.20, i.e. less than a standard deviation, the logarithm of the ratio of standard deviations is not far from zero (0.174), and the coverage proportion is above 0.90 for both treatment groups.

A negative but negligible effect of a larger than one number of siblings on the outcome is estimated, which amounts to −0.042 (s.e. = 0.027).

The central role of family background and structure is also pointed out when we are interested in evaluating the effects of different factors on the educational attainments of pupils, such as comparing province versus city schools (T_2). Indeed, by ignoring the family features, improperly comparing the success rates of pupils enrolled into the schools of Bologna with those of the province, we derive a positive and significant effect, which amounts to 0.036 (s.e. = 0.017). Conversely, controlling for the family characteristics leads to a negative and negligible effect (−0.013, s.e. = 0.024). We compare 593 pupils students enrolled into a school of the city of Bologna with 602 matched units enrolled into a province school, after assessing the overlap in the covariate distribution ($\hat{\Delta}_{01} = 0.46$, logarithm of the ratio of standard deviations = 0.130, $\hat{\pi}_0 = 0.91$, $\hat{\pi}_1 = 0.95$).

Table 2 Estimation results

Causal factor	Treated	Controls	$\hat{\Delta}_{01}$	Log ratio sd	Overlap $\hat{\pi}_1$	Overlap $\hat{\pi}_0$	$\widehat{ATT}$ (sd)
T_1	280	943	0.20	0.174	0.90	0.97	−0.042 (0.027)
T_2	593	602	0.46	0.130	0.91	0.95	−0.013 (0.024)

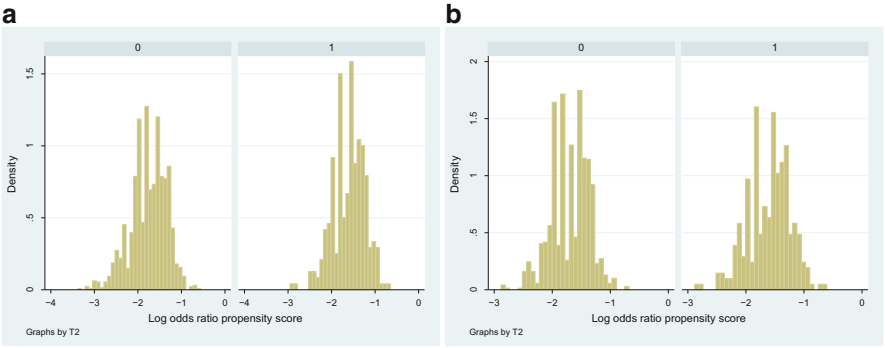

Fig. 1 Histogram of log odds ratio propensity score for treated and controls of T_1. (**a**) Before matching. (**b**) After matching

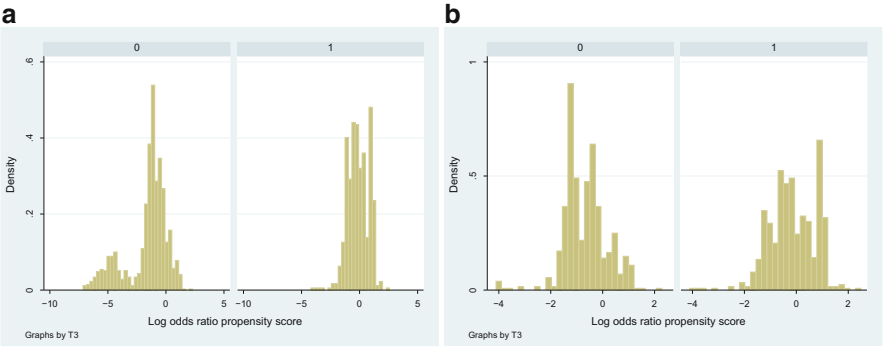

Fig. 2 Histogram of log odds ratio propensity score for treated and controls of T_1. (**a**) Before matching. (**b**) After matching

Finally, Figs. 1 and 2 graphically reveal considerable imbalance between the two treatment arms for both T_1 and T_2 before matching, and the reached balance after matching strategy. We consider the log odds ratio transformation of propensity scores (Imbens and Rubin 2014).

Concluding Remarks

We provided a statistical framework toward a proper investigation of the factors which may influence the achievement of students in the different levels of their education in observational studies on individual learning experiences and, more generally, in evaluation researches of educational field. We adopted a causal inference perspective, employing a matching strategy based on propensity score to evaluate the effects on success/failure at the end of the first upper secondary school year of two factors under study: the number of siblings and the school location.

The role of family background and structure is shown to be fundamental as either causal factors or characteristics to be controlled for. Thus, collecting more accurate information on family (such as educational and job qualification of parents, income and marital status, as well as students features) becomes even more important, as it ensures a valid framework to investigate different aspects of the phenomenon and, as a final result, to properly define local educational policies.

References

Becker, S.O., Ichino, A.: Estimation of average treatment effects based on the propensity scores. Stata J. **2**, 358–377 (2002)

Erikson, R., Goldthorpe, J.H., Jackson, M., Yaish, M., Cox, D.R.: On class differentials in educational attainment. Proc. Natl. Acad. Sci. **102**(27), 9730–9733 (2005)

Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, and the Social and Biomedical Sciences. Cambridge University Press, Cambridge/New York (2014)

Imbens, G.W., Wooldridge, J.M.: Recent developments in the econometrics of program evaluation. J. Econ. Lit. **47**(1), 5–86 (2009)

Lauer, C.: Family background, cohort and education: a French-German comparison based on a multivariate ordered probit model of educational attainment. Lab. Econ. **10**, 231–251 (2003)

Mignani, S., Monari, P., Stracqualursi, L.: A categorical data model to assess critical points. Presented at the meeting "Innovazione e Società", Firenze (2011)

Mignani, S., Pillati, M., Martelli, I.: Un indicatore statistico del background familiare nello studio del successo scolastico degli studenti della provincia di Bologna. Sociol. Lav. **120**(IV), 194–214 (2011)

Rosenbaum, P., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)

Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)

Part VI

New Methodological Developments in Social and Demographic Studies

Developing a Composite Indicator of Residents' Well-Being: The Case of the Romagna Area

Cristina Bernini, Andrea Guizzardi, and Giovanni Angelini

Abstract

There is a growing literature on the assessment of quality of life conditions and well-being in geographical or administrative areas. The paper proposes a new measure of Subjective Well Being (SWB) based on residents' satisfaction with Personal life domain, Leisure activities, and satisfaction with Life as a whole. The SWB index is constructed by using a Composite Indicator, where the weights are calculated by means of common weight DEA-like models. Results evidence that in an economically developed area as the Romagna, Leisure activities is the domain most affecting SWB. Differences of SWB across sub-areas and among different socio-demographic groups of residents are also detected.

Introduction

Researchers have used various approaches to define and measure the complex and multidimensional construct of quality of life (QOL), such as social indicators, subjective well being measures (SWB), and economic indices. In these frameworks, Composite Indicators (CIs) have increasingly been accepted as a useful tool for benchmarking, performance comparisons, policy analysis, and public communication (Zhou et al. 2010). A CI is a mathematical aggregation of a set of sub-indicators for measuring multi-dimension concepts that cannot be captured by a single indicator. The way individual indicators are weighted and aggregated is the main critical aspect of this approach, affecting directly the quality and reliability

C. Bernini (✉) • A. Guizzardi • G. Angelini
Department of Statistical Sciences, University of Bologna, Bologna, Italy
e-mail: cristina.bernini@unibo.it; andrea.guizzardi@unibo.it; giovanni.angelini3@unibo.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_29,

of the resulting CIs (Saisana et al. 2005). To overcome subjectivity in determining the weights, in this paper we suggest using an optimization approach able to look for endogenous weights for each indicator considered. We propose to determine the weights for sub-indicators by solving a data envelopment analysis (DEA) linear programming problem that requires no *prior* knowledge of the weights for sub-indicators. In particular, the common weight approach is used, being useful to evaluate the relative importance of different domains in affecting SWB.

We apply the proposed approach to measure the residents' SWB of various groups of people who differ by gender, age, and administrative sub-area where they live. Following Diener et al. (1999), we define SWB as a broad construct that includes people's satisfaction with Personal life domains, Leisure activities, and satisfaction with Life as a whole (Overall or General life satisfaction). As a case study, we consider residents in the Romagna area and in the close State of S. Marino (Italy), a developed territory homogeneous in term of economic and social standards, and cultural heritage.

This paper furthers the existing literature in several ways. Firstly, a DEA approach in constructing the SWB index based on the subjective evaluations of individuals is used. At our knowledge no empirical studied have estimated community SWB by using DEA models. Secondly, a common weight approach is implemented, allowing to evaluate the relative influence of the different domains on the SWB index. Finally, the proposed approach is applied to a novel sample of residents in the Romagna area.

Theoretical Framework

In the last decades the interest on the measurement of community quality of life (QOF) and SWB has been increased, being the living conditions of an area of primary concern of public policies. A field of empirical literature measures SWB directly by using the subjective evaluation of a person's quality of life (Diener 1984; Diener et al. 1999; Kahneman and Krueger 2006), considering individuals' subjective experience of their lives and subjective perceptions of their social environment. The underlying assumption is that for understanding the well-being of an individual, it is important to directly measure the individual's cognitive and affective reactions to her or his whole life, as well as to specific domains of life (for a review see Diener 1984; Diener et al. 1999).

In developing SWB measures, the bottom-up spillover theory has reached a large consensus (Sirgy et al. 2008, 2010). The bottom-up theory (Andrews and Withey 1976; Campbell et al. 1976) asserts that life satisfaction is influenced by satisfaction with several life domains (satisfaction with community, family, work, social life, health, and so on). Satisfaction with a particular life domain is, in turn, influenced by lower levels of life concerns within that domain. The bottom-up spillover theory

postulates that satisfaction within a specific life domain accumulates and vertically spills over to super-ordinate domains (for a review see: Diener 1984; Diener et al. 1999).

Following this stream of research, the paper aims to develop a new indicator of subjective measure of community well-being based on bottom-up spillover theory. Even if in the literature there is a general agreement that SWB is a composite of satisfaction with a number of domains in life, there is little agreement on which domains actually constitute QOL and should thus be included in SWB measures. In the study, we propose to construct a novel measure of SWB by considering three main domains affecting individual's well-being and quality of life, namely: Personal life satisfaction, Leisure satisfaction, and General life satisfaction.

The first domain considered in the analysis regards Personal life. Following Diener (1984), Diener et al. (1999), and Sirgy et al. (2010), a SWB measure is partially determinate by the satisfaction that individuals assign to their personal life. This domain is related to several aspect of individual life, as health, work, marriage and family, physical fitness, income, standard of living, neighborhood, etc.

Satisfaction with Leisure activities is also recognized to affect individual's SWB, because leisure provides opportunities to meet life values and needs (Diener et al. 1999). Through participation in leisure activities people build social relationships, feel positive emotions, acquire additional skills and knowledge, and therefore improve their quality of life (Rodriguez et al. 2008; Brajša-Žganec et al. 2011; Dolnicar et al. 2011).

Satisfaction toward General life is also considered in the analysis. The domain captures the overall satisfaction on one's life in general. Measures of overall life satisfaction are frequently reported in the literature of quality of life studies and have much established validity (Andrews and Withey 1976; Campbell et al. 1976; Sirgy et al. 2010).

The Method: CI indicators and Common Weight DEA-Like Model

Composite Indicators (CIs) have increasingly been accepted as a useful tool for benchmarking, performance comparisons, policy analysis, and public communication (Zhou et al. 2010). A CI is a mathematical aggregation of a set of sub-indicators for measuring multi-dimension concepts that cannot be captured by a single indicator. Its construction, however, is not straightforward and involves a number of steps that need to be carefully examined. After choosing the model for the multidimensional concept that is being measured, individual indicators have to be measured and a *weighting and aggregation* technique should be defined. Following Saisana et al. (2005), the last step is a major one, affecting directly the quality and reliability of the resulting CIs. The Authors underline that the determination of weights could take advantage of additional information from experts but, whatever

the weighting and aggregation technique, it may be difficult to reach an agreement on such weights among the entities compared as each entity has its own specificity.

Subjectivity in determining the weights for sub-indicators is the main problem in the construction of a CI. To overcome this problem, in this paper we suggest using an optimization approach able to look for endogenous weights. We propose to determine the weights for sub-indicators by solving a DEA linear programming problem that requires no prior knowledge of the weights for sub-indicators. In order to rank all the DMUs on the same scale, a common weights solution is adopted (Despotis 2002; Zohrehbandian et al. 2010). The approach not only differentiates efficient DMUs but also estimates the relative weight of each factor (in our case the different domains) in the SWB indicator.

Formally, assume that there are n entities (that are individuals), whose CI*s* are to be calculated based on m sub-indicators. Let I_{ij} denote the value of entity j with respect to sub-indicator i. Without loss of generality, all the sub-indicators are assumed larger than unity and of the benefit type (they satisfy the property of "the larger the better"). The purpose is to aggregate I_{ij} ($i = 1, 2, \ldots, m$) into a composite indicator $\mathrm{CI}_{j,}$ for entity j by using the weight w_i for sub-indicator i:

$$\mathrm{CI}_j = \sum_{i=1}^{m} w_i I_{ij}, \quad j = 1, 2, \ldots, n \tag{1}$$

We suggest using DEA model to determine the weights, DEA is a lineal programming technique traditionally used to estimate the efficiency of a Decision Making Unit (DMU) within production contexts characterized by multiple outputs and inputs. The DEA technique has come to be used in a broader number of applications, not only "productive," evidencing the flexibility of DEA as a valid tool of multidimensional analysis. In recent years, a novel use of DEA has been in the analysis of standards of living and social well-being (Zhou et al. 2010) and territorial diffusion index of infrastructures (Mazziotta and Vidoli 2009).

In particular, in this paper a DEA-like model with common weights approach is implemented. Differently from the traditional DEA model, in the DEA-like model all sub-indicators are viewed as "output" because they are assumed of the benefit type, while a single "dummy input" with value unity is assigned to each DMU. The common weights approach assumes that the weighting structure is the one maximizing the score of each entity subject to the constrain that it has to be unique across entities. This approach has several appealing features. It allows either to differentiate efficient DMUs or to estimate the relative influence of each factor (in our case the different domains) on the SWB indicator. Therefore, the common weights approach reveals to be useful for the governance of a territory interested in evaluating changes of residents' evaluations across domains.

As pointed in Despotis (2002), a general formulation to obtain a common set of weights is as follows:

$$\min \quad t \cdot \frac{1}{n} \cdot \sum_{j=1}^{n} d_j + (1-t) \cdot z$$

s.t.

$$\begin{aligned} &\sum_{i=1}^{m} I_{ij} w_i + d_j = E_j^*, \quad j = 1, \ldots, n \\ &d_j - z \leq 0, \quad j = 1, \ldots, n \\ &w_i \geq \varepsilon, \quad i = 1, \ldots, m \\ &d_j \geq 0, \quad j = 1, \ldots, n \\ &z \geq 0 \end{aligned} \tag{2}$$

For $t = 1$, the objective function to be minimized represents the arithmetic mean of d_j, that is the difference between the score obtain using a common set of weights and the optimum score E_j^* obtain with the classical DEA model without the constraint of the common weights. The choice is consistent with the theoretical goal of achieving a ranking that represents a collective optimal, where the structure of weights (common weights) ensures the maximum sum (or mean) of CIs.

The Data

Data were collected in the Romagna area in the period January–March 2010, by conducting a telephone survey. The sampling design was based on stratification with respect to: provinces (Rimini, Forlì and Cesena, Ravenna, San Marino) and demographic characteristics (age and gender). The final sample consists of 810 questionnaires. The detailed characteristics of the sample are summarized in Table 1.

Following the SWB approach, the questionnaire is constructed to capture residents' satisfaction in respect to three main dimensions that are: satisfaction with Personal life domains, Leisure activities, and Life as a whole. In particular, we suggest evaluating the three different domains by using the following items:

1. Personal life domain: material status, health, work, family, religion/spirituality.
2. Leisure activities: social relationships, sport activities, hobby, shopping, culture, entertainment, holiday.
3. General life satisfaction: with personal life, with the main life dimensions, with respect to personal goals, compared with peers.

Residents were required to give a score (using a Likert-scale 1–7) to each item related to the different dimensions of well-being, expressed in terms of satisfaction. The mean values of items, classified in respect to domains, are reported in Table 2.

A preliminary exploratory factor analysis is performed in order to individuate the factors underlying the items of the questionnaire and their internal consistency. In Table 3 the results of the analysis are presented. We can observe that the loadings associated with almost all the items are quite high, indicating that the items are influenced significantly by the corresponding underlying construct. Furthermore, all

Table 1 Profile of respondents

		N	Percentage
Provinces	Forlì-Cesena	252	31.1
	Ravenna	248	30.6
	Rimini	250	30.9
	San Marino	60	7.4
Gender	Male	421	52.0
	Female	389	48.0
Age	<25	65	8.0
	25–35	119	14.7
	35–45	178	22.0
	45–55	130	16.0
	55–65	96	11.9
	>65	222	27.4

Table 2 Item description and mean values

Personal life		Leisure activities		General life	
Item	Mean	Item	Mean	Item	Mean
Material status	4.21	Social relationships	4.70	Personal life	4.73
Health	4.67	Sport activities	3.91	Main life dimensions	4.68
Family	4.74	Hobby	4.53	Personal goals	4.68
Friends	4.67	Culture	4.40	Compared with peers	4.60
Work	4.24	Entertainment	4.26		
Spirituality/religion	3.96	Shopping	4.36		
		Holiday	4.60		

Table 3 Exploratory factor analysis

Personal life		Leisure activities		General life	
Item	Factor loadings	Item	Factor loadings	Item	Factor loadings
Material status	0.64	Social relationships	0.61	Personal life	0.74
Health	0.75	Sport activities	0.55	Main life dimensions	0.83
Family	0.79	Hobby	0.71	Personal goals	0.78
Friends	0.76	Culture	0.69	Compared with peers	0.72
Work	0.56	Entertainment	0.73		
Spirituality/religion	0.10	Shopping	0.59		
		Holiday	0.65		
Cronbach's alpha		Cronbach's alpha		Cronbach's alpha	
0.75 (without spirituality/ religion 0.82)		0.83		0.85	

the factors show a Cronbach's alpha higher than 0.75 being quite good (Nunnally 1970). The index is slightly low only for the Personal life domain, but its value improves when spirituality/religion is excluded (0.82) from the construct.

Results

DEA Model Estimates

In the analysis we used the three constructs obtained by the factor analysis (Personal life domain, Leisure activities, and General life), having a higher validity. The factor weights are calculated by using a DEA-like model, maximizing CI as defined in (2), s.t.

$$\mathrm{CI}_j = \sum_{i:1}^{3} w_i I_{ij} \leq 1{,}000, \quad j = 1, 2, \ldots, n, \quad 50 \leq w_i \leq 1{,}000$$

Estimated coefficients w_i (Table 4) represent the weights that maximized the overall satisfaction, that is the maximum sum of satisfaction scores (CI_j) of the interviewed residents. The estimated weights evidence that Leisure activities is the most affecting factor in SWB of residents in Romagna, the General life the lowest one. As in Brajša-Žganec et al. (2011), we find empirical evidence to sustain the general hypothesis of a strong positive relationship between participation in leisure and SWB. People through participation in leisure activities improve their Quality of Life. Personal life domain also contributes to SWB. The result is consistent with Sirgy et al. (2008): higher satisfaction with family, friend, and health reflects in an increase of SWB.

SWB Score Distributions

Having calculated the DEA-like common weights, a composite SWB index for each resident may be obtained by Eq. (1). Then, individual indicators are averaged with respect to different geographical and socio-demographical partitions, in order to investigate the existence of significant differences among residents' clusters in the Romagna area. The SWB indicators evidence some interesting results (Table 5).

Firstly, residents in San Marino show a higher SWB than residents in the Romagna area. The result is expected in the sense it is consistent with the diffused feeling of a very high well-being in the independent state of San Marino. Only residents in Rimini show a SWB score not significantly different from that observed in San Marino. A spillover effect of SWB is also detected: moving away from San Marino, SWB decreases progressively, reaching the lowest scores in the province of Ravenna. Secondly, SWB decreases with age and labor market position. As in Brajša-Žganec et al. (2011), older and retired residents show lower value of the SWB index than younger residents. Having time for leisure activities does not necessarily imply a greater SWB, as demonstrated by the gap between the two groups with the largest and lowest amount of free time: students (young people) and retired (old people). Conversely, the role of an active planning over the life time is confirmed by the fact that the SWB is higher for people who have a better

Table 4 Weight estimates by common weight DEA-like model

Domains	Weights
Personal life	347.05
Leisure activities	426.50
General life	294.15

Table 5 Subjective well-being indexes

Provinces	CI	Age	CI	Labor market position	CI
Forlì-Cesena	0.61	<25	0.72	Self-employed/manager	0.67
Ravenna	0.58	25–35	0.64	White collar/teacher	0.65
Rimini	0.67	35–45	0.68	Blue collar	0.60
San Marino	0.71	45–55	0.60	Retired	0.58
		55–65	0.61	Student	0.71
		≥65	0.58	Other	0.63
Whole sample 0.63					

chance of personal realization (self-employed and manager), respect to those who are engaged in routine tasks, such as blue collars or retired people. Finally we find three groups that are significantly different in terms of SWB scores: self-employed, manager and white collar; blue collars and retired people; and students.

Conclusions

We developed a new measure of SWB based on residents' evaluation of three different life domains, namely Personal life, Leisure activities, and Global life. The domains have been defined by using the well established bottom-up spillover theory of life satisfaction. The composite index approach is followed to obtain SWB measures. DEA-like model with common weights is used to construct the SWB index. This approach is particularly effective in evaluating public policies, because DEA weights may be seen as simple and objective measures of the relative influence of residents' life domains. Data on residents' satisfaction were collected in the Romagna area in 2010 and support the construct validity of the measures proposed in the analysis.

Results reveal that Leisure activities is the domain most affecting the SWB index in the Romagna area. Thus, policy makers should pay much attention to develop policies and programs directed to the enhancement of residents' leisure life. The domain is the one where public governments have the larger opportunity to make direct interventions. The development of programs and services related to community parks, recreation, sports, are typical examples.

SWB distributions with respect to different socio-demographic characteristics evidence that young people and students are the most satisfied of their well-being. The result is consistent with a huge literature showing that SWB decreases with age. SWB also changes with respect to the different administrative areas, being the

highest in the independent State of S. Marino. Surprisingly, the SWB decreases as the geographical distance from San Marino increases, suggesting a possible spillover effect that needs to be further investigated.

References

Andrews, F.M., Withey, S.B.: Social Indicators of Well-Being. Plenum, New York (1976)
Brajša-Žganec, A., Merkaš, M., Šverko, I.: Quality of life and leisure activities: how do leisure activities contribute to subjective well-being? Soc. Indicators Res. **102**, 81–91 (2011)
Campbell, A., Converse, P.E., Rodgers, W.J.: The Quality of American Life: Perceptions, Evaluations, and Satisfaction. Russell Sage, New York (1976)
Despotis, D.K.: Improving the discriminating power of DEA: focus on globally efficient units. J. Oper. Res. Soc. **53**, 314–323 (2002)
Diener, E.: Subjective well being. Psychol. Bull. **95**(3), 542–575 (1984)
Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L.: Subjective well-being: three decades of progress. Psychol. Bull. **125**(2), 276–302 (1999)
Dolnicar, S., Yanamandram, V., Cliff, C.: The contribution of vacations to quality of life. Ann. Tourism Res. **39**(1), 59–83 (2012)
Kahneman, D., Krueger, A.: Developments in the measurement of subjective well being. J. Econ. Perspect. **20**, 3–24 (2006)
Mazziotta, C., Vidoli, F.: La costruzione di un indicatore sintetico ponderato. Un'applicazione della procedura Benefit of Doubt al caso della dotazione infrastrutturale in Italia. Ital. J. Reg. Sci. **8**(1), 35–69 (2009)
Nunnally, J.C.: Introduction to psychological measurement. McGraw-Hill, New York (1970)
Rodriguez, A., Latkova', P., Sun, Y.: The relationship between leisure and life satisfaction: application of activity and need theory. Soc. Indicators Res. **86**, 163–175 (2008)
Saisana, M., Saltelli, A., Tarantola, S.: Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. J. Roy. Stat. Soc., Ser. A (Stat. Soc.) **168**(2), 307–323 (2005)
Sirgy, M.J., Gao, T., Young, R.F.: How residents' satisfaction with community services influence quality of life (QOL) outcomes? Appl. Res. Qual. Life **3**(2), 81–106 (2008)
Sirgy, M.J., Widgery, R., Lee, D., Yu, G.: Developing a measure of community well-being based on perceptions of impact in various life domains? Soc. Indicators Res. **96**, 295–311 (2010)
Zhou, P., Ang, B.W., Zhou, D.Q.: Weighting and aggregation in composite indicator. construction: a multiplicative optimization approach. Soc. Indicators Res. **96**, 169–181 (2010)
Zohrehbandian, M., Makui, A., Alinezhad, A.: A compromise solution approach for finding common weights in DEA: an improvement to Kao and Hung's approach. J. Oper. Res. Soc. **61**, 604–610 (2010)

New Technologies and Statistics: Partners for Environmental Monitoring and City Sensing

Rina Camporese, Giovanni Borga, Niccolò Iandelli, and Antonella Ragnoli

Abstract

Urban space is interconnected, thanks to a vast array of technological devices whose data can be aggregated in a geographic database thereby providing a representation of what is happening around us. City Sensing is an "immersive sensing" and a new opportunity to survey the territory and the environment, by means of low-cost sensors small enough to be wearable. Advantages of such a framework are the widespread and numerous measurements at lower unit cost and also the near real-time friendly communication, together with an interaction with citizens. There are, of course, some limits: low-cost sensors' measurements are affected by a greater error; the huge amount of data produced can result in a sort of data overload; pressure for real time can lead to hasty elaborations. Statistics can offer some help to reduce the impact of the drawbacks related to measurement quality control and error estimates, and they can also offer possible solutions for significant data synthesis and representation.

R. Camporese (✉)
Italian National Institute of Statistics, New Technologies and Information Territory and Environment, Iuav University of Venice, Tolentini Venezia, Italy
e-mail: rina.camporese@gmail.com

G. Borga • N. Iandelli • A. Ragnoli
New Technologies and Information Territory and Environment, Iuav University of Venice, Tolentini Venezia, Italy
e-mail: giovanni@borga.it; niccogeo@gmail.com; aragnoli@libero.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_30,

City Sensing and New Technologies

A new strategy for environmental monitoring is outlined by the rapid development of sensors and computer networks: a great number of data acquisition instruments, distributed and interconnected, provide near real-time data flows.

Recent technological research has produced sensors—mainly based on Micro-Electro-Mechanical System—that can be integrated into commonly used instruments (i.e. smartphones or devices small enough to be wearable and at low cost). These sensors can either measure various environmental quantities by translating variations of physical parameters into electrical impulses (e.g. acceleration, temperature, humidity, concentration of gases, magnetic fields, . . .), or transform built-in microphone in a noise detector. Some sensors are: already installed in common existing instruments and can be activated with user friendly software applications (e.g. smartphones can be used as noise sensors); easily integrated into daily life instruments, thanks to plug-and-play tools (e.g. temperature and humidity tool for cell phones); inserted in stand-alone small instruments, connected to the web (e.g. wearable air pollution sensors streaming real-time data on the internet) (Fig. 1).

Global Navigation Satellite Systems (GNSS) have entered people's everyday life and pockets (Hofmann-Wellenhof et al. 2008), and each mobile phone could become an environmental station and a node of a larger monitoring network (Calabrese et al. 2007).

A wide spread monitoring network could displace the traditional paradigm of environmental monitoring based on the use of few stand-alone stations, drawing the attention to the pervasiveness of low-cost nodes, equipped with light sensors meant to get a small gridded representation of the territory. As a result, a detailed spatial representation of environmental phenomena could be obtained.

Urban space could be, in some cases it already is, interconnected, thanks to an impressive range of technological devices whose data can be aggregated in a geographic database, providing a relevant representation of what is happening around us. Having this in mind, City Sensing becomes an immersive sensing and a new exciting opportunity to survey the territory (Fig. 2).

In combination with the Web 2.0 opportunities, City Sensing can be defined as a Sensor Web, to perform environmental monitoring in the style of social networking and from a cooperative perspective (Calabrese et al. 2009). Several social actors and stakeholders can use common technological instruments to gather environmental data, share them on collaborative web platforms, in the aim of obtaining a shared knowledge of environmental status.

In 2004 Goodchild and Janelle edited an inspiring book entitled *Thinking spatially in the social science* where they elaborated the idea that space and territory are essential to study social—and environmental—phenomena; space and territory also constitute the basis for integrating data from different sources, thanks to geographical coordinates.

Fig. 1 Smartphone app to measure noise pollution (NoiseTube), wind sensor for smartphone, wearable air pollution sensor (Sensaris)

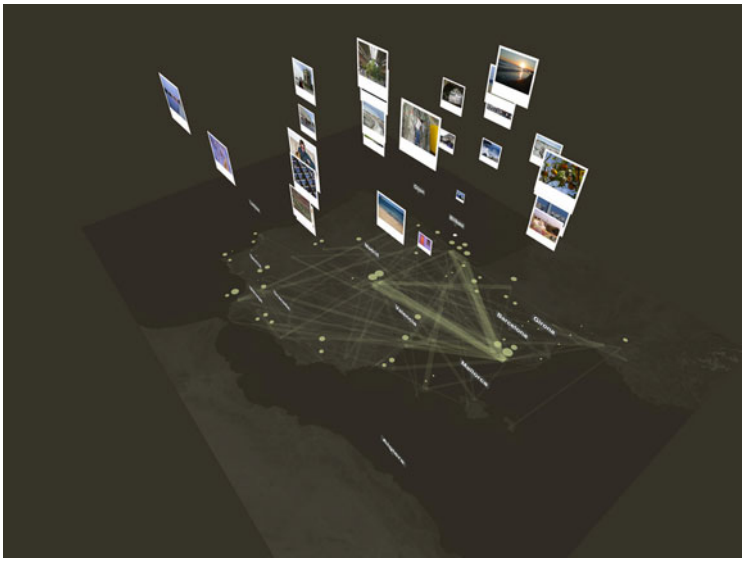

Fig. 2 Los Ojos del Mundo—the World's Eyes (courtesy of Senseable City Lab, MIT, http://senseable.mit.edu/worldseyes/visuals.html)

Nowadays, such spatial approach to social and environmental sciences can be quite easily applied, thanks to Geographic Information Science (Burrough 2001; Longley et al. 2010) and, in particular, as a result of data mash-ups technologies (Batty et al. 2010).

Accessible web geographic technologies become, therefore, a means to integrate data coming from disparate sources, a way to create synergic information and a platform to share and communicate knowledge.

Potentials, Drawbacks and Possible Solutions

The main advantages of such a framework are the widespread and numerous measurements at lower unit cost, the near real-time friendly communication and the possible interaction with citizens. On the contrary, until recently, environmental measurements have been extremely rare both in time and space, since they were only obtained, thanks to very precise and expensive means.

There are, of course, some limits to this new approach. Firstly, data coming from actual low-cost sensors are affected by a greater error as compared to certified expensive instruments. Secondly, a huge amount of data can be easily and quickly produced; this can result in a sort of data overload, which is difficult to manage and interpret. Furthermore, the pressure for real-time data can lead to hasty and un-meditated elaborations. Not to mention privacy issues.

Statistics can offer some help to limit these drawbacks with regard to measurement quality control and error estimates (Goodchild 2008).

The main advantages of using a statistical approach could be:

- Rationalise the numerous and enthusiastic data collection processes, so as to make them more significant and representative, e.g. in terms of sampling strategies
- Raise awareness of measurements' quality control and evaluation of errors
- Keep into consideration the uncertainty of the results
- Enhance the essential role of metadata
- Expertise in statistical disclosure control, with regard to privacy issues.

In cooperation with Information Design, statistics can also develop innovative solutions in favour of a significant data synthesis and representation (Tufte 2005, 2007), especially when multidimensional data have to be considered along with both space and time.

The spread of these scenarios has opened the door to new research experiences made by the Iuav NT&ITA New Technologies and Information on Territory and Environment Research Group, such as the design of an integrated system of sensors for environmental and road traffic monitoring (widespread in the territory and based on WSN—Wireless Sensor Network), the test of a prototype wearable multi-sensor with Bluetooth transmission, the evaluation of data quality for hand-held sensors of gas concentration and urban noise, and a research project on the possible integration of institutional and citizens' knowledge on noise pollution.

Here are two examples of how new technologies can modify the traditional approach to environmental monitoring. The Framework for the Development of Environment Statistics defined by the United Nations, which is currently under revision, has been taken into consideration as a reference frame for the following reflections (United Nations 1984, 1988, 1991).

Air Pollution

As to air pollution, UNSD Environmental Indicators essentially take only emissions into consideration, while indicators on ambient concentrations of selected pollutants are not present, mainly because they lack quality, coverage across countries and international comparability.

Spatial patterns of air pollutants concentration vary significantly across territories and they are usually monitored with very few stations. Furthermore, data on environmental pollution are not always open to the public. As a result, sometimes national environmental statistics describe the characteristics of the monitoring network (i.e. air monitoring stations: number, type and location), instead of pollutants concentration (i.e. the environmental measures obtained by the monitoring network); this is the case of Italy, for instance.

The following quotation comes from a UN document dated 1991: "The cost of environmental monitoring has inhibited the development of statistically valid space/time sampling frames" (United Nations 1991); it clearly explains the reason behind the state of the art.

Low-cost sensor networks open a new scenario, where challenges are no more related to the costs of measurement, but to instruments calibration, proper time and space dependent sample strategies, ascertainment of statistical validity, and significant data reduction of massive data sets.

There are several examples of experimental projects on air quality assessment based on participative contributions of citizens equipped with low-cost, portable sensors. To mention a few as an example:

- CamMobSens: Cambridge University Mobile Urban Sensing is an air pollution monitoring project using data coming from pedestrians and cyclists equipped with hand-held sensors together with data coming from more traditional monitoring fixed units (Fig. 3) http://www.escience.cam.ac.uk/mobiledata/
- Copenhagen Wheel: a project of Senseable City Lab of MIT University in Boston to use ordinary bicycles to map pollution levels, traffic congestion and road conditions in real-time http://senseable.mit.edu/copenhagenwheel/

Tests carried out at Iuav University, using metal oxide semi-conductor gas sensors, have showed that such a perspective is promising but some steps have still to be made to improve the quality of measurements obtained from low-cost sensors, especially with regard to calibration for temperature and humidity.

Noise Pollution

In 1988 UN selected a suitable indicator for noise pollution: the population exposed to excessive noise, i.e. noise levels exceeding national standard (United Nations 1988). Furthermore, the 2002 EU Directive on Environmental Noise required Member States to draw harmonised strategic noise maps (European Parliament and Council 2002).

Fig. 3 Real-time pollution monitoring using hand-held sensors carried by pedestrians (courtesy of Cambridge Mobile Urban Sensing, http://www.escience.cam.ac.uk/mobiledata/)

Despite that, actual national statistics on noise pollution often show only the responses to noise pollution, in terms of actions and policies adopted to reduce noise pollution effects, which are essentially the Responses described in the DPSIR framework—Driving forces, Pressures, State, Impacts, Responses—defined by OCSE and European Environment Agency (European Environment Agency 1999).

In Italy, for example, noise barriers, low noise road surfaces and noise zoning are the selected indicators on "noise pollution" in "Urban Environment Indicators" statistical national report (Istat 2009).

Iuav University of Venice and Veneto Region Environmental Agency are currently carrying out a research project in order to evaluate the quality of noise measurements coming from mobile phones applications (free and commercial), as compared to the ones derived from professional noise metres. The work is in progress, but the preliminary results are promising: despite some loss in measurement quality, data seem acceptable in the aim of obtaining generalised maps of urban noise pollution.

At the time being, low-cost noise sensors appear to guarantee a better performance in terms of measurement quality (provided that they are calibrated), as compared to sensors measuring concentration of gases, for which the output measurements are more controversial.

Therefore, a hypothetical sample strategy to assess environmental noise in Italy is proposed below. It aims at obtaining noise exposure maps along the roads in urban environments, using two indicators quoted in the EU Directive: a day-evening-night level in decibels and a night-time noise indicator (obtained through A-weighted

long-term average sound levels, determined over all the day periods of a year) (European Parliament and Council 2002).

The proposed sample strategy requires stratification according to space and time. As to space, road segments of urban environment could be stratified by techno-functional characteristics related to speed limits and traffic flow (highways, suburban, urban, local). Such information is available in the Catasto Strade (Roads Register), required by law and usually available, in some form, at least for principal towns. Another spatial stratification variable could be the land cover class, such as the one provided by the GSE Land European Urban Atlas Services (part of the European Earth Observation Programme—GMES). It comes from a very high-resolution hot spot mapping of urban functional areas and it allows for their stratification according to different urban fabric density (continuous, dense, medium, low, sparse) and to functional characteristics (residential, industrial, etc.).

As to time stratification, the sample strategy could resemble the one adopted for HETUS—Harmonised European Time Use Survey, which covers an entire 12 months period—24 h—7 days, with stratification based on month and type of day (Mon–Fry, Sat, Sun) (Eurostat 2009).

The characteristics of small noise sensors in terms of cost and transportability would easily adapt to such a sample. If a noise map has to be the output, estimates of noise indicators derived from sampled locations would then be used as expected values for the road segments that have not been surveyed, on the basis of spatial stratification variables.

Wiki Monitoring

In this view there comes an unconventional proposal: to contaminate the traditional sampling approach with a wiki component, in the style of collaborative mapping—OpenStreetMap—and collaborative research—GalaxyZoo (Fig. 4).

The first experience is well known and it shows how Web 2.0 collaborative activities can produce a valid map of the territory. The second is probably less known: a data set made up of a million galaxies' images collected by the robotic telescope of the Sloan Digital Sky Survey have been made available on the web and the morphological classifications of galaxies, which enables scientists to understand how galaxies form and evolve, is being carried out by a network of registered web users, after a brief online tutorial phase (Keel et al. 2011). This latter experience shows how common citizens with an interest in astronomy are open to follow simple guidelines to contribute to a scientific project, in the aim of creating a wide knowledge framework.

Another emblematic experience in this field is NoiseTube (Fig. 5): a mobile phone application developed for a research project of the Vrije Universiteit of Brussel, which fosters a new participative approach to noise pollution monitoring by involving the general public (Maisonneuve et al. 2010; D'Hondt and Stevens 2011).

Fig. 4 Galaxy Zoo: morphological classification of galaxies carried out by trained web users (courtesy of Galaxy Zoo Project, http://www.galaxyzoo.org/)

Fig. 5 NoiseTube: noise pollution monitoring using mobile phones (courtesy of Vrije Universiteit Brussel BrusSense group, http://www.brussense.be/)

Traditional environmental measurements could then be integrated with spontaneous contributions of citizens who wear portable instruments and become themselves sensors (Goodchild 2007), capturing data with smartphone applications;

the participative contribution would cover areas and periods of time which cannot be covered by the institutional and traditional survey fieldwork.

The final estimates would be produced through ex-post weight calibration and proper weighted averages of both structured and wiki components of the sample.

Furthermore, mobile devices applications usually allow to integrate quantitative measurement of a phenomenon (e.g. decibel for noise pollution) and individual perception and opinion of people (comments, tags, . . .).

The comparison of both objective measures on the state of the environment and subjective perceptions of that same environment by people who live there generate new knowledge, analogously of what happens in health statistics when measured and perceived health is combined.

With this regard, a reference experience is represented by EyeOnEarth.eu site of the European Environment Agency that publishes air and water monitoring data coming from national agencies on a web portal which is open to the comments and evaluations of general public (Jirícek and Di Massimo 2011).

This goes on the same direction as the recent evolution of Quality of Urban Life studies, which is making efforts to integrate the two historical approaches: one focusing on objective measurements and the other paying attention to subjective evaluations and appraisals (Marans and Stimson 2011).

A Test for Noise Pollution in Padova: An Inspiration for General Reflections

Despite the fact that the described scenario seems to foster a good potential in the authors' opinion, it is not free from problems and drawbacks.

One of the major concerns is related to the possible bias caused by the digital divide. Furthermore, such a change in the technical instruments and in data collection procedures produces a totally new set of sources for non-probabilistic errors which has yet to be dealt with. But, one could say "what survey doesn't suffer from fieldwork errors?" and keep on working to evaluate them and find out solutions to them.

Authors have worked to integrate institutional urban noise data coming from the local environmental agency (Arpav) and spontaneous contributions from mobile phone applications, in the aim of verifying the possibility to produce a joint urban noise map along the routes of the city of Padova.

Environmental monitoring via technological instruments is not for everybody. Some technical skills are required together with the willingness to respect certain rules in the aim of obtaining correct measurements. For example noise pollution measures along the streets should be performed according to a set of few, but strict, rules, ranging from the way the phone is held in hand to the minimum duration of the measured time slot. Therefore, the participants had to be trained, and both technical and motivational aspects of the project had to be transmitted to them.

When people are involved, strategies to protect data confidentiality are essential because, in order to obtain this kind of data, it is necessary to know the detailed

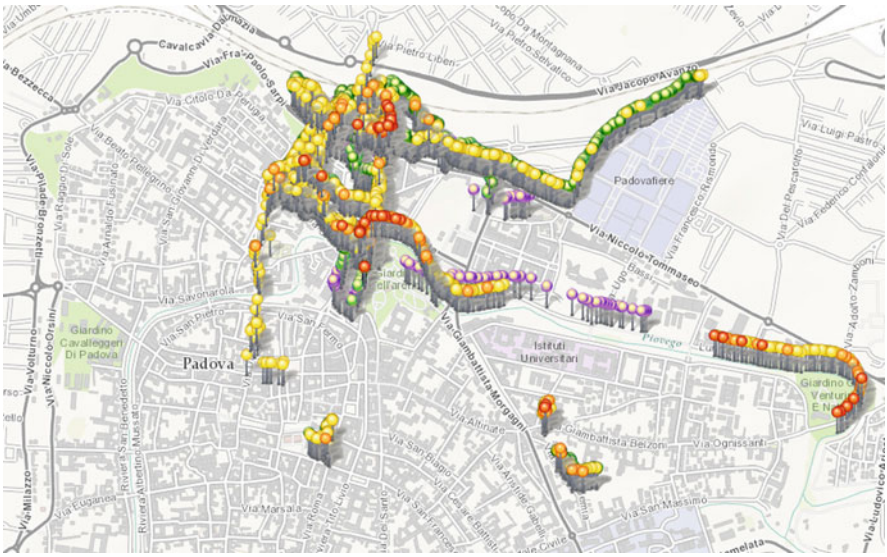

Fig. 6 Test noise map of Padova. Short Link: (bit.ly/wBYgIG)

geographical locations of participants to the project, which is a very sensible piece of information. Some methods to protect privacy are easy to imagine (different views related to zoom levels, buffers of proper size to mask identities, ad hoc coordinates' transformations to prevent overlapping on common reference systems), others have to be invented. Identificative and personal information of contributors should only be used during the analysis phase: results show to the public must be aggregated or mask the identity of contributors.

Ethical and methodological problems arise, too, due to the fact that technologies go extremely into details with regard to people's life and interfere with people's behaviour (Fisher 2011).

While facing such problems, the authors' view is based on the idea that technology reveals what the human being is. The way technologies are used mainly depends on the intentions and emotions of the users. This can be considered true also when users are not individuals, but groups or institutions. Therefore, risks of misuse are always at the door and have to be taken into account, together with positive potentials.

As far as data disclosure and confidentiality are concerned, an intriguing perspective is adopted by the European Union Joint Situation Centre for spatial tools dissemination. When deciding whether to disclose spatial data upon request, the evaluation process takes into account the risks deriving from the particular transaction acted on data, instead of evaluating data themselves. Transaction includes professional skill of personnel, processing methods and published results. Thereafter, non-risky transactions on potentially harmful data are allowed (Claeys et al. 2011). This approach sounds more sensible than the most common one which

simply closes the access to potentially harmful data, no matter what they are used for.

With regard to data disclosure control, the risk of omitting relevant knowledge processes that are not performed because of the fear of contraindications should be taken in account, too.

Some preliminary results obtained with NoiseTube application to measure noise level with smartphones have been published on the web (Fig. 6).

On the whole, the task requires a multidisciplinary approach, involving acoustic, geographic information science, web technologies, information and communication expertise, social science, etc. (not in order of importance).

It is complicated, but challenging, too.

References

Batty, M., Crooks, A., Hudson-Smith, A., Milton, R., Anand, S., Jackson, M., Morley, J.: Data Mash-Ups and the Future of Mapping. JISC, Bristol (2010)

Burrough, P.A.: GIS and geostatistics: essential partners for spatial analysis. Environ. Ecol. Stat. **8**, 361–377 (2001)

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-Time Urban Monitoring Using Cellular Phones: A Case-Study in Rome. MIT Press, Boston (2007)

Calabrese, F., Kloeckl, K., Ratti, C.: WikiCity: real-time location-sensitive tools for the city. In: Foth, M. (ed.) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, Information Science Reference. Hershey, New York (2009)

Claeys, C.: Ensuring security whilst keeping access. Principles of security for the European space tools. In: International Conference on Data Flow from Space to Earth, Venice, 21–23 March (2011)

European Parliament and Council: Directive 2002/49/EC on Assessment and Management of Environmental Noise. Official Journal of the European Communities (18.7.2002)

D'Hondt, E., Stevens, M.: Participatory noise mapping. In: Adjunct Proceedings of the 9th International Conference of Pervasive Computing, pp. 33–36, June, 2011

European Environment Agency: Environmental indicators: Typology and overview. Technical report No. 25 (1999)

Eurostat, European Commission: Harmonises European time use surveys 2008 guidelines. Methodologies and working papers (2009)

Fisher, K.: Narrative mediated by gadgets: ethical and methodological implications. In: 33rd International Association for Time Use Research Conference, Oxford, 1–3 August, 2011

Goodchild, M.F.: Citizens as sensors: web 2.0 and the volunteering of geographic information (Editorial). In: GeoFocus International Review of Geographical Information Science and Technology, vol. 7, pp. 8–10 (2007)

Goodchild, M.F.: Statistical perspectives on geographic information science. Geographical Anal. **40**, 310–325 (2008)

Goodchild, M.F., Janelle, D.G. (eds.): Spatially Integrated Social Science. Oxford University Press, New York (2004)

Hofmann-Wellenhof, B., Lichtenegger, H., Wasle, E.: GNSS – Global Navigation Satellite Systems. GLONASS, Galileo, and + more. Wein, New York (2008)

Istat: Indicatori Ambientali Urbani. Istituto Nazionale di Statistica, Roma (2009)

Jirícek, Z., Di Massimo, F.: Microsoft Open Government Data Initiative (OGDI), eye on earth case study. In: Hrebícek, J., Schimak, G., Denzer, R. (eds.) Environmental Software Systems. Frameworks of eEnvironment. IFIP Advances in Information and Communication Technology, vol. 359, pp. 26–32. Springer, Boston (2011)

Keel, W.C., Chojnowski, S.D., Bennert, V.N., Schawinski, K., Lintott C.J., Lynn, S., Pancoast, A., Harris, C., Nierenberg, A.M., Sonnonfeld, A., Proctor, R.: The Galaxy Zoo survey for giant AGN-ionized clouds: past and present black-hole accretion events. Mon. Not. R. Astron. Soc. **420**, 868–900 (2012)

Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W.: Geographic Information Systems and Science, 3rd edn. Wiley, New York (2010)

Maisonneuve, N., Stevens, M., Ochab, B.: Participatory noise pollution monitoring using mobile phones. Inf. Polity **15**(1–2), 51–71 (2010)

Marans, R.W., Stimson, R.J. (eds.): Investigating Quality of Urban Life. Springer, New York (2011)

Tufte, E.R.: Visual Explanations. Images and Quantities, Evidence and Narrative. Graphic Press, Cheshire (2005)

Tufte, E.R.: The Visual Display of Quantitative Information, 2nd edn. Graphic Press, Cheshire (2007)

United Nations: A Framework for the Development of Environment Statistics. Statistical Papers, Series M No. 78, United Nations, New York (1984)

United Nations: Concepts and Methods of Environment Statistics: Human Settlements Statistics? A Technical Report. Studies in Methods, Series F No. 51, United Nations, New York (1988)

United Nations: Concepts and Methods of Environment Statistics – Statistics of the Natural Environment. Studies in Methods, Series F No. 57, United Nations, New York (1991)

Recent Developments in Multidimensional Analysis for Customer Satisfaction

Luigi D'Ambra and Enrico Ciavolino

Abstract

The paper aims at showing some recent methodological evolutions for analyzing Customer Satisfaction (CS) models, by using Structural Equation Model (SEM) based on the information theoretic approach, and the method of the Cumulative Correspondence Analysis (CCA) in case of ordered categorical variables in a multifactor state system based on Taguchi's statistic.

Introduction

Knowledge of CS is an advantage for companies operating in a competitive market where the ability to listen and maximizing CS, can be a key for the success, and how to steer the company towards a system of organizational excellence. In order to depict CS, the researcher has to deal with both quantitative and qualitative data. In the quantitative data, the most used approach is represented by the Structural Equation Models (SEMs), where the CS and its components are measured by latent variables, through a set of manifest variables. In case of qualitative data, some recent applications have shown the potentiality of the Correspondence Analysis (CA) and its generalization to the ordinal data, in modelling and predicting the choices of the customer and the level of CS, in transport public service Sarnacchiaro and D'Ambra (2011).

L. D'Ambra (✉)
Dipartimento di Economia, Management, Istituzioni Università di Napoli Federico II, Naples, Italy
e-mail: dambra@unina.it

E. Ciavolino
History, Society and Human Studies, Università del Salento, Lecce, Italy
e-mail: enrico.ciavolino@unisalento.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_31,

In this chapter we will go through a review and synthesis of some recent methodological developments in the field of SEMs and CA applied to the passenger transport and the job satisfaction.

The remainder of the paper is organized as follows. Section 1 introduces the Generalized Maximum Entropy (GME) for the SEMs and some recent developments. Section 2 illustrates the relationship between the Correspondence Analysis and the Taguchi's statistic. Section 3 shows the generalization of CA, called Cumulative Correspondence Analysis. In the last section are reported some brief conclusions.

The GME Structural Equation Modeling

It is well known in literature that the estimation methods for analyzing data for SEM in a *soft* and in a *hard* way are, respectively, Partial Least Squares (Wold 1966, 1982) and Maximum Likelihood Estimation (Jöreskog 1973; Bollen 1989). A middle way between the soft and hard modelling is represented by the GME estimation method.

The GME estimation approach (Golan et al. 1996; Golan 2008) is based on the Shannon's Entropy (1948). The GME approach for the SEM (Ciavolino and Al-Nasser 2009) considers the *Re-Parameterization* of the unknown parameters and the disturbance terms as a convex combination *of expected value of a discrete random variable*. The coefficient matrices of classical equations system of SEM (Bollen 1989), $\mathbf{B}$, $\mathbf{\Gamma}$, $\mathbf{\Lambda}^{y}$, $\mathbf{\Lambda}^{x}$, and the co-variance matrices, $\mathbf{\Phi}$, $\mathbf{\Psi}$, $\mathbf{\Theta}^{\varepsilon}$, $\mathbf{\Theta}^{\delta}$ are all re-parameterized as expected values of discrete random variable with M fixed points for the coefficients and N for the errors. The three SEM equations can be re-formulated as a unique function model which represents a consistency constraint:

$$\mathbf{y}_{(p,1)}=\mathbf{\Lambda}^{y}_{(p,m)}\cdot\left(\mathbf{I}_{(m,m)}-\mathbf{B}_{(m,m)}\right)^{-1}\cdot\left\{\mathbf{\Gamma}_{(m,n)}\cdot\mathbf{\Lambda}^{x-1}_{(n,q)}\cdot\left(\mathbf{x}_{(q,1)}-\boldsymbol{\delta}_{(q,1)}\right)+\boldsymbol{\tau}_{(m,1)}\right\}+\boldsymbol{\varepsilon}_{(p,1)}$$

With m endogenous latent variables, n exogenous latent variables and p and q manifest endogenous and exogenous variables. Given the re-parameterization and the re-formulation, the GME system can be expressed as a constrained non-linear programming problem. The coefficients and the error terms are estimated by recovering the probability distribution of the discrete random variables set. The vectors: $\mathbf{p}^{\mathrm{B}}$, $\mathbf{p}^{\Gamma}$, $\mathbf{p}^{\Lambda^{y}}$, $\mathbf{p}^{\Lambda^{x}}$, $\mathbf{w}^{\tau}$, $\mathbf{w}^{\varepsilon}$, $\mathbf{w}^{\delta}$ are calculated by the maximization of the following entropy function:

$$\begin{aligned}
&H\left(\mathbf{p}^{\mathrm{B}},\mathbf{p}^{\Gamma},\mathbf{p}^{\Lambda^{y}},\mathbf{p}^{\Lambda^{x}},\mathbf{w}^{\tau},\mathbf{w}^{\varepsilon},\mathbf{w}^{\delta}\right)=\\
&\quad-\mathbf{p}^{\mathrm{B}'}_{(1,m\cdot m\cdot M)}\cdot\ln\mathbf{p}^{\mathrm{B}}_{(m\cdot m\cdot M,1)}-\mathbf{p}^{\Gamma'}_{(1,m\cdot n\cdot M)}\cdot\ln\mathbf{p}^{\Gamma}_{(m\cdot n\cdot M,1)}-\mathbf{p}^{\Lambda^{y}{}'}_{(1,p\cdot n\cdot M)}\cdot\ln\mathbf{p}^{\Lambda^{y}}_{(p\cdot m\cdot M,1)}\\
&\quad-\mathbf{p}^{\Lambda^{x}{}'}_{(1,q\cdot n\cdot M)}\cdot\ln\mathbf{p}^{\Lambda^{x}}_{(q\cdot n\cdot M,1)}+\\
&\quad-\mathbf{w}^{\tau'}_{(1,m\cdot N)}\cdot\ln\mathbf{w}^{\tau}_{(m\cdot N,1)}-\mathbf{w}^{\varepsilon'}_{(1,p\cdot N)}\cdot\ln\mathbf{w}^{\varepsilon}_{(p\cdot N,1)}-\mathbf{w}^{\delta'}_{(1,q\cdot N)}\cdot\ln\mathbf{w}^{\delta}_{(q\cdot N,1)}
\end{aligned}$$

subjected to the *consistency* and *normalization constraints* (the sum of each coefficients and the error terms probability vector have to be equal to 1).

The main advantages of using GME estimation method are its desirable properties (Golan et al. 1996; Golan 2008), which can be briefly summarized in the following points:

- The GME approach uses all the data points and does not require restrictive moments or distributional error assumptions.
- Thus, unlike other estimators, the GME is robust for a general class of error distributions.
- The GME estimator may be used when the sample is small, where there are many covariates, and when the covariates are highly correlated.
- Moreover, using the GME method, it is easy to impose nonlinear and inequality constraints.

In the following sub-paragraphs we report some theoretical extension of the SEM–GME, for the analysis of same specific customer satisfaction models.

Spatial-SEM for the Passenger Satisfaction

In the transport sector, not only the particular feature of the transport sector might influence the measurement of the passenger satisfaction, but also by some spatial effects attributable to the territorial dislocation of the stations (Gallo and Ciavolino 2009).

In this context a Spatial Structural Equation Models (S-SEM) specification has been introduced as a flexible tool for modeling multivariate spatial data and answering research questions about latent factors underlying spatial samples data. The GME formulation for the S-SEM which assumes a fixed effect spatial lag specification has been derived to deal with the problems of endogeneity and collinearity implied by the introduction of spatial effects.

Since we are observing N different units (passengers) located in different positions (stations), we have to take into account the effects that the geographic position can generate into the model. Spatial structures are generally associated to: (a) *Absolute location* effects which are relevant to evaluate for each observation the impact of being located at a particular point in space, and to (b) *Relative location* effects that consider relevant the position of a observation relative to other observations. The first effect called *spatial heterogeneity* assumes that each observation can have its own characteristic for the phenomenon under investigation. Moreover, in the latter case, it is assumed that the value observed in a sample in a specific location h can be affected by the value observed in another location k, with $h \neq k$. This effect, called *spatial dependence*, is due to the spatial interaction between contiguous observations.

The *spatial unobserved heterogeneity among* spatial observations is allowed by introducing fixed effects in the measurement model (Bernardini Papalia 2006).

For the *spatial dependence*, we focus on one of the widely used approach (called *spatial LAG model*) where the spatial correlation pertains to the dependent variable. In this context, it is assumed interdependence of latent variables across areas. This assumption may be formalized by including a set of spatial lag variables into the

measurement model, which represents the relationship between the manifest and latent variables. In doing this, a spatial weights matrix $\mathbf{W}$ of non-stochastic time constant weights has to be specified. This is a $(N \times N)$ matrix in which the rows and columns correspond to the cross-sectional locations (Bernardini Papalia and Ciavolino 2011).

More specifically, using the measurement exogenous model equation, the set of latent exogenous variables $\boldsymbol{\xi}$ is enlarged to include: (1) *Spatial Lag variable*, that is the first-order contiguity spatially lagged dependent variable, here considered as exogenous and so defined with X instead of Y; (2) The *fixed effects*, that are year and country dummies; (3) and the set of q exogenous manifest variables $\mathbf{X}_{NT,q}$.

$$\begin{aligned}\mathbf{w}^x &= \text{Spatial-Lag} = (\mathbf{I}_{T,T} \otimes \mathbf{W}_{N,N}) \cdot \mathbf{x}_{NT,1} \\ \mathbf{d}^t &= \text{Dummy-Times} = (\mathbf{I}_{T,T} \otimes \mathbf{1}_{N,1}) \\ \mathbf{d}^s &= \text{Dummy-Space} = (\mathbf{1}_{T,1} \otimes \mathbf{I}_{N,N)})\end{aligned}$$

To take into account also the spatial lag variable and both the fixed effects, the exogenous measurement model is rewritten as follow:

$$\mathbf{X}^*_{NT,q+1+T+N} = \left[\mathbf{X}_{NT,q} \left| \left(\mathbf{I}_{T,T} \right) \otimes \mathbf{W}_{N,N} \right) \cdot \mathbf{x}_{NT,1} \left| \left(\mathbf{I}_{T,T} \right) \otimes \mathbf{1}_{N,1} \right) \right| \left(\mathbf{1}_{T,1} \otimes \mathbf{I}_{N,N} \right) \right]$$

The above equation reports the specification of the units in the measurement model, reporting the vector $\mathbf{x}_{q,1}$ of the q manifest exogenous variables in the form of matrix $\mathbf{X}_{NT,q}$, where the variables are reported in the columns and the units in the rows.

Then, the associated $\boldsymbol{\Lambda}^x$ matrix which specifies the regression coefficients of the observed variables on the latent exogenous variables, is defined as $\boldsymbol{\Lambda}^x = [\boldsymbol{\tau}|\rho|\boldsymbol{\alpha}]$, including: the set of the *exogenous variables coefficients* ($\boldsymbol{\tau}$), the *spatial autoregressive parameter* (ρ), the vector of *fixed effects*, relative to *time and spatial effects* defined as $\boldsymbol{\alpha} = [\alpha_t|\alpha_s]$.

The matrices $\mathbf{I}_{T,T}$, $\mathbf{I}_{N,N}$, $\mathbf{1}_{T,1}$, and $\mathbf{1}_{N,1}$ are respectively the identity matrices and the vectors of one for the panel of N units within T periods. The symbol $\otimes$ is the Kronecker product. The matrix formulation of the exogenous measurement model can be reformulated considering the spatial and the fixed effects, by placing $k = T + N$, as below reported:

$$\mathbf{x}_{q+1+k,1(t)} = \boldsymbol{\Delta}^x_{q+1+k,l(t)} \cdot \boldsymbol{\xi}_{l,1(t)} + \boldsymbol{\delta}_{q+1+k,1(t)}$$

The exogenous measurement model is extended in this way adding to the q manifest exogenous variables, the spatial lag variable, the time and the spatial effect, that means $1 + k$ rows.

Multi-Group SEM for the Job Satisfaction

A commonly encountered situation in the analysis if the job satisfaction is the study of workers coming from different populations; in these situations it results important to understand how the populations might differ. Some examples of job satisfaction analyzed as multi-group SEM are referred to test the invariance of the measurement models for workers in private and in public sector (Ciavolino 2009) or to test the invariance in the job satisfaction structural coefficients, for workers with different levels of educations (Carpita et al. 2012).

The extension of the SEM–GME for the inclusion of the groups effect is obtained by the definition of the following fit function, which is a weighted combination of the above-defined entropy fit function:

$$H_G(\mathbf{p}, \mathbf{w}) = \sum_{g=1}^{G} \frac{N_g}{N} \times f_g\left[H\left(\mathbf{p}_g, \mathbf{w}_g\right)\right]$$

where N_g is the sample size in the gth group, $N = N_1 + N_2 + \cdots + N_G$,$f_g[\cdot]$ is the fit function for the gth group, and $\mathbf{p}_g$ and $\mathbf{w}_g$ are probabilities vectors for the coefficients and error terms, relative to each group.

Comparability test is made by deciding which elements or matrices of parameters should be tested for equality across groups and in which order these tests should be made, the choice depends on the theory behind the models or on the empirical needs.

Once the hypotheses have been tested, the fit of the more constrained model can be compared with that of the previous hypothesis with the chi-square difference test statistic (Likelihood Ratio). If the fit of the constrained model is much worse than that of the less restricted model, it is possible to conclude that only the first hypothesis has to be considered true.

In a multi-group SEM the measures of model fit are analogous to those in single group analyses. To test the hypothesis, the GME provides a normalized entropy measure that quantifies the level of information in the data, giving a global measure of the goodness of relationships. The normalized entropy measure (Golan 1996) can be expressed by the following formulation:

$$S(\mathbf{p}) = \left(-\mathbf{p}' \cdot \ln \mathbf{p}\right) / K \cdot \ln M\Big)$$

This normalized index is a measure of the reduction in uncertainty information, where—$\mathbf{p}'\ln(\mathbf{p})$ is the Shannon's entropy function defined only the structural coefficients; K is the number parameters ($\mathbf{B}$, $\mathbf{\Gamma}\,\mathbf{\Lambda}^{y}$, and $\mathbf{\Lambda}^{x}$) to estimate and M is the number of the Fixed Points.

If the fit of the constrained model is much worse than that of the less restricted model, it is possible to conclude that only the first hypothesis has to be considered not refused.

Analysis of Ordered Categorical Data

Correspondence Analysis and the Taguchi's Statistic

The correspondence analysis of a two-way contingency table is a popular statistical tool for graphically identifying the nature of the association between the categorical variables that form the table. It has been used by the data analysts from a variety of disciplines over the past 50 years; however, little attention has been paid to the case where the variables are ordinal. Of course, there have been some contributions that deal with ordered categorical variables, including those of Parsa and Smith (1993), Ritov and Gilula (1993), and Schriever (1983). Generally these procedures involve constraining the output obtained from applying singular value decomposition (SVD) so that the coordinates, or the components of SVD, have an ordered structure. An alternative approach involves using the moment decomposition (MD) procedure of Beh (1997). However, all of these procedures involve partitioning the Pearson chi-squared statistic using the elements of the SVD or the MD. Another approach performing the correspondence analysis when the cross-classified variables have an ordered structure is to consider the Taguchi's statistic (Taguchi 1974). The Taguchi's statistic takes into account the presence of an ordinal categorical variable by considering the cumulative sum of cell frequencies across the variable. Therefore one may consider the impact of differences between adjacent ordered categories on the association between the row and column categories.

Measure of the Association Proposed by Taguchi Cumulative Chi-Squared Statistic

In a single factor experiment, let A is a factor with I levels and an equal number, n, of observations taken at each level. Suppose that an experiment was conducted and the discrete outcomes were classified into one of K ordered categories. Let n_{ik} denotes the observed frequency in category k at the ith level of the factor ($k = 1, \ldots, K$; $i = 1, \ldots, I$) of a two-way contingency table, $\mathbf{N}$, that cross-classifies n individuals/units according to I row categories and K ordered column categories. Let $n_{i.}$ and $n_{.k}$ be the ith row and kth column marginal frequencies, respectively. Also let $p_i = \frac{n_i}{n}$ be the (i, i)th element of the diagonal matrix $\mathbf{D}_I$ and $p_{.k} = \frac{n_{.k}}{n}$. The cumulative frequencies are determined by $Z_{ik} = \sum_{j=1}^{k} n_{ij}$. Similarly denote $d_k = \sum_{j=1}^{k} n_j / n = \sum_{j=1}^{k} p_j$ as the cumulative relative frequency up to kth column category.

Sometimes one has experimental data that needs to be collapsed to be useful to the multifactor setting. More specifically, let say, we have more than one factors in an $I \times K$ contingency table. Assume that I is the total number of experiments

conducted in an orthogonal array and observations taken at each run on a K ordered categorical response. To obtain a collapsed table Taguchi determines the factor effects. For example, the effect of the first level of factor A (i.e., A1) is determined by the experiments conducted at that level and then sum the observations in each cumulative category. In this paper we will use this collapsed data structure to discuss our proposed methodology and find the best factor levels for quality improvement.

The following statistic, which is proposed by Taguchi (1974), measures the association between the rows and the column variables:

$$T = \sum_{k=1}^{K-1} w_k \left[\sum_{i}^{I} n_i \left(\frac{Z_{ik}}{n_i} - d_k \right)^2 \right],$$

where $w_k = [d_k(1-d_k)]^{-1}$ or $w_k = 1/(K-k)$

Under the hypothesis of independence, $w_k = [d_k(1-d_k)]^{-1}$ is proportional to the inverse of the conditional expectation of the kth term. Unlike the Pearson chi-squared statistic, the Taguchi's T is finitely bounded by the interval $[0, n(I-1)]$.

Suppose we let $\mathbf{y}_i = (n_{i1}, \ldots, n_{iK})$' be as a vector of observed frequencies for the ith row and let $\mathbf{W}$ be the $(K-1)\times(K-1)$ diagonal matrix of the weights w_k, for $k = 1, \ldots, K-1$. Nair (1987) showed that the Taguchi's statistic T can be expressed in matrix notation by:

$$T = \sum_{i=1}^{I} \left(\mathbf{y}_i \mathbf{A}' \mathbf{W} \mathbf{y}_i' \right) / n_i .$$

where $\mathbf{A}$ is the $(K-1)\times K$ matrix involving the cumulative column relative marginal frequencies d_k such that:

$$\mathbf{A} = \begin{pmatrix} 1-d_1 & -d_1 & -d_1 & -d_1 \\ 1-d_2 & 1-d_2 & -d_2 & -d_2 \\ \cdots & \cdots & \cdots & \cdots \\ 1-d_{K-1} & 1-d_{K-1} & 1-d_{K-1} & 1-d_{K-1} \end{pmatrix}.$$

Nair (1987) also demonstrated that the link between the Pearson chi-squared statistic and the Taguchi's statistic is

$$T = \sum_{K=1}^{K-1} X_k^2.$$

Here X_k^2 is the Pearson chi-squared for $I\times 2$ contingency table obtained by aggregating column categories 1 to k and aggregating the column categories $(k+1)$ to K. For this reason T is also referred to as cumulative chi-squared statistic.

Cumulative Correspondence Analysis Using the Taguchi's Statistic

Theoretical Aspects

Generally, correspondence analysis can be performed by partitioning the Pearson chi-squared statistic into the sum of squares of the singular values of the Pearson ratio's. Here we consider instead the Taguchi's statistic proposed by Beh et al. (2011) and D'Ambra et al. (2009). An interesting application on the CS (transport public service) can be found in Sarnacchiaro and D'Ambra (2011).

If the user wishes to identify how the ordered column categories impact upon the association between the two categorical variables, consider again the Taguchi's statistic, the statistic may be expressed in full matrix notation by:

$$T = n \cdot \text{trace}\left(\mathbf{D}_I^{-1/2}\mathbf{NA}'\mathbf{WND}'\mathbf{D}_I^{-1/2}\right)$$

where the element of the diagonal matrix $\mathbf{D}_I^{-1/2}$ is the $1/\sqrt{p_i}$ and $p_i = n_i/n$. Therefore, the inertia for Taguchi's statistics will be:

$$T/n = \text{trace}\left(\mathbf{D}_I^{-1/2}\mathbf{NA}'\mathbf{WND}'\mathbf{D}_I^{-1/2}\right)$$

and this is referred to as the Taguchi's inertia. The matrix $\mathbf{A}'\mathbf{WA}$ could be decomposed such that $\mathbf{A}^{\mathrm{T}}\mathbf{WA} = \mathbf{Q}\boldsymbol{\Gamma}\mathbf{Q}'$ where, $\mathbf{Q}$ is matrix containing the singular vector (with general term q_{km}), $\boldsymbol{\Gamma}$ is the diagonal matrix containing the eigenvalues γ_k where $(m = 1, \ldots, M)$ and $M = \min(I, K-1)$.

$$\sum_{k=1}^{K-1} \gamma_k q_{km} q_{km'} = \begin{cases} 1 \text{ if } m = m' \\ 1 \text{ if } m \neq m' \end{cases}$$

Therefore if $\mathbf{u}_i = u_{i1}, \ldots u_{iK-1} = \mathbf{Q}'\mathbf{y}_i/\sqrt{p_i}$ then Taguchi's inertia can be expressed as:

$$\frac{T}{n} = \sum_{k=1}^{K-1} \gamma_k \left(\sum_{i=1}^{I} u_{ik}^2\right) = \text{trace}\left(\mathbf{U}'\Gamma\mathbf{U}\right)$$

where γ_k is the kth largest eigenvalue of $\mathbf{A}'\mathbf{WA}$ and the (k, k)th element of the diagonal matrix $\boldsymbol{\Gamma}$. Suppose we let:

$$\mathbf{Y} = \Gamma^{1/2}\mathbf{U}' = \mathbf{W}^{1/2}\mathbf{AN}'\mathbf{D}_{\mathbf{I}}^{-1/2}$$

then the Taguchi's inertia T/n may be expressed as:

$$\frac{T}{n} = \text{trace}\left(\mathbf{Y}'\mathbf{Y}\right)$$

In terms of the Taguchi's statistic, we can perform SVD on $\mathbf{Y}$ such that $\mathbf{Y} = \tilde{\mathbf{A}}\mathbf{D}_\lambda\mathbf{B}'$. Here $\tilde{\mathbf{A}}$ is a $(K-1)\times M$ matrix of singular vectors associated with the difference between each pair of adjacent column categories and $\tilde{\mathbf{B}}$ is a $(I \times M)$ matrix of singular vectors for I row categories. For both matrices $M = \min(I, K-1)$ and they are subject to the constraint $\tilde{\mathbf{A}}'\tilde{\mathbf{A}} = \mathbf{I}$ and $\tilde{\mathbf{B}}'\tilde{\mathbf{B}} = \mathbf{I}$. The diagonal elements of $\mathbf{D}_\lambda$ are real and positive and are the first $M = \min(I, K-1)$ singular values of $\mathbf{Y}$. They are arranged in descending order so that $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_M$. By considering the SVD of $\mathbf{Y}$, it can be verified using the properties $\tilde{\mathbf{A}}'\tilde{\mathbf{A}} = \mathbf{I}$ and $\tilde{\mathbf{B}}'\tilde{\mathbf{B}} = \mathbf{I}$ that the mth singular value can be expressed as:

$$\lambda_m = \sum_{i=1}^{I}\sum_{k=1}^{K-1}\sqrt{\frac{w_k}{p_{i.}}\tilde{a}_{km}\tilde{Z}_{ik}\tilde{b}_{im}}$$

where a_{km} is the (k, m)th element of $\tilde{\mathbf{A}}$, b_{im} is the (i, m)th element of $\tilde{\mathbf{B}}$ and $\tilde{Z}_{ik} = Z_{ik}/n$. Also, Taguchi's inertia may be expressed as the sum of squares of these singular values since $T/n = \text{trace}(\mathbf{Y}'\mathbf{Y}) = \text{trace}(D_\lambda^2)$. We note for $I > 2$ and in the case of equiprobable categories the eigenvectors are given by chebychev polynomials So, the first component (linear) is proportional to the Kruskal-Wallis statistic for contingency tables. The second component (quadratic) is the generalization of the grouped data version of Mood's statistic.

Graphical Representation

In order to visualize the association between the ordered column categories and the nominal row categories the set columns of $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{A}}$ may be considered. That is, the coordinate of the row category in M-dimensional space may be defined by $\{b_{im} : m = 1, \ldots M\}$ while the variation between the kth pair of adjacent, and ordered, column categories can be visualized by considering the coordinate (b_{km}: $m = 1, L, M$) for $k = 1, L, K-1$.

These coordinates are analogous to standard coordinates used in the classical correspondence analysis (Greenacre 1984, p. 93) and do not reflect the nature of the association between the two variables. Instead, one may consider the following row and column coordinates:

$$F = D_I^{-1/2}\tilde{\mathbf{B}}\mathbf{D}_\lambda \tag{1}$$

$$\mathbf{G} = \mathbf{W}^{-1/2}\tilde{\mathbf{A}}\mathbf{D}_\lambda \tag{2}$$

respectively. By considering such scaling, the Taguchi's inertia can be expressed as:

$$T/n = \text{trace}\left(\mathbf{F}^T\mathbf{D}_I\mathbf{F}\right) = \text{trace}\left(\mathbf{G}^T\mathbf{W}\mathbf{G}\right) = \text{trace}\left(\mathbf{D}_\lambda^2\right)$$

so that the metric of the nominal row coordinates is $\mathbf{D}_I$ and the metric of the column categories is $\mathbf{W}$. By defining the row coordinates by (2) the variation of each row category can be graphically depicted as a low dimensional subspace. Similarly, since the Taguchi's statistic takes into consideration the ordered column categories by determining the nature of the variation between each adjacent category, the coordinates $\mathbf{G}$ describe the variation between each adjacent category. Therefore, any variation between each adjacent column category can be determined simultaneously with each row category. For correspondence analysis this is a unique interpretation of the plot. One must note that for the column coordinates $\mathbf{G}$ the position does not reflect the position of each ordered column category, but does reflect the impact of each pair of adjacent column categories on the association between the two variables. Other properties see Beh et al. (2011). A generalization of this approach concerning the double cumulative correspondence analysis was presented in International Conference "Correspondence Analysis and Related Methods"—CARME 2011 (Rennes 2011) and submit to international journal. In this paper are reported the links with classical correspondence analysis and proposed a unified approach.

Conclusions

In this paper we presented a theoretical review of some recent developments for the analysis of quantitative and categorical data for the evaluation of the passenger and the job satisfaction. It is showed how the methods SEM–GME and CCA can help the researcher in the interpretation of the CS in case of spatial, multi-group, and ordinal data.

Our intention is not to give a detailed examination of these topics, but to present how these methods can be useful for the handling of this kind of data in the framework of the customer satisfaction analysis.

References

Beh, E.J.: Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. Biom. J. **39**, 589–613 (1997)

Beh, E., D'Ambra, L., Simonetti, B.: Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic. Commun. Stat. **40**(9), 1620–1632 (2011)

Bernardini Papalia, R.: Modeling Mixed Spatial Processes and Spatio-Temporal Dynamics in Information Theoretic Frameworks. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT 2006, pp. 1483–1491. Physica Verlag Heidelberg, New York (2006)

Bernardini Papalia, R., Ciavolino, E.: GME estimation of spatial structural equations models. J. Classif. **28**(1), 126–141 (2011)

Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)

Carme: International conference on Correspondence Analysis and Related Methods Correspondence Analysis and Related Methods – Rennes, 8–11 February 2011

Carpita, M., Ciavolino, E.: Using the GME estimator with the rasch analysis in multigroup SEM. Quad. di Stat. **14**, 61–64 (2012)
Ciavolino, E., Al-Nasser, A.D.: Comparing generalized maximum entropy and partial least squares methods for structural equation models. J. Nonparametric Stat. **21**(8), 1017–1036 (2009)
Ciavolino, E.: Job Satisfaction in Private and Public Sector: An information theoretic model for the Multi-group Analysis. In: INNOVAZIONE E SOCIETÀ: Metodi e politiche per la valutazione dei servizi, Facoltà di Economia, Università degli Studi di Brescia. 24–26 Giugno (2009)
D'Ambra, L., Koksoy, O., Simonetti, B.: Cumulative correspondence analysis of ordered categorical data from industrial experiments. J. Appl. Stat. **36**(12), 1315–1328 (2009)
Gallo, M., Ciavolino, E.: Multivariate statistical approaches for the customer satisfaction into transportation sector. Global Local Econ. Rev. **13**(2), 55–70 (2009)
Golan, A.: Information and entropy econometrics—a review and synthesis. Found. Trends Econ. **2**(1–2), 1–145 (2008)
Golan, A., Judge, G., Miller, D.: Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, New York (1996)
Greenacre, M.J.: Theory and Application of Correspondence Analysis. Academic, London (1984)
Jöreskog, K.G.: A General Method of Estimating a Linear Structural Equation System. In: Goldberg, S.A., Duncan, D.O. (eds.) Structural Equation Models in the Social Sciences, pp. 85–112. Seminar Press, New York (1973)
Nair, V.N.: Chi-squared type tests for ordered alternatives in contingency tables. J. Am. Stat. Assoc. **82**, 283–291 (1987)
Parsa, A.R., Smith, B.S.: Scoring under ordered constraints in contingency tables. Commun. Stat. – Theory Methods **22**, 3537–3551 (1993)
Ritov, Y., Gilula, Z.: Analysis of contingency tables by correspondence models subject to ordered constraints. J. Am. Stat. Assoc. **88**, 1380–1387 (1993)
Sarnacchiaro, P., D'Ambra, A.: Cumulative correspondence analysis to improve the public train transport. Elec. J. Appl. Statist. Anal.: Deci. Supp. Syst. and Serv. Eval. **2**(1), 15–24 (2011)
Schriever, B.F.: Scaling of order dependent categorical variables with correspondence analysis. Int. Stat. Rev. **51**, 225–238 (1983)
Shannon, C.E.: A mathematical Theory of Communications. Bell Syst. Tech. J. **27**, 379–423 (1948)
Taguchi, G.: A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. Saishin Igaku **29**, 806–813 (1974)
Wold, H.: Estimation of Principal Components and Related Models by Iterative Least Squares. In: Krishnajah, P.R. (ed.) Multivariate Analysis, pp. 391–420. Academic, New York (1966)
Wold, H.: Soft Modelling: The Basic Design and Some Extensions. In: Jöreskog, K.G., Wold, H. (eds.) Systems Under Indirect Observations, Part 2, pp. 1–54. North-Holland, Amsterdam (1982)

Tourism Statistics for Destination Management: The Trips/Arrivals Model

Stefano De Cantis and Mauro Ferrante

Abstract

Simply counting the number of tourists in a destination is not as simple as one might initially think. The present work intends to formalize a conceptual model by decomposing tourism nights, trips, and average duration of visit in a given destination, in terms of quantities and parameters that only for a small part could be derived from official tourism statistics. On the other hand, most of the quantities and parameters required for determining the number of tourists in a given destination are in general unknown, and need to be (directly or indirectly) estimated. Empirical evidences resulting from a survey on incoming tourism in Sicily are provided, showing the biases affecting supply-side statistics on tourism when they are used as a measure of tourism flows in a destination. The changing nature of demand and the increasing segmentation of the holyday market also raise the need for more accurate, destination-based, information, and the model proposed can assist destination managers and researchers in facing the problem of quantifying and analysing tourism behaviours.

The paper is a common work of both the authors; however, Sects. 2, 2.1, 2.3, and 4 are due to Stefano De Cantis, and Sects. 1, 2.2, and 3 are due to Mauro Ferrante.

S. De Cantis (✉) • M. Ferrante
Dipartimento di Scienze Economiche, Aziendali e Statistiche (SEAS), Università degli studi di Palermo, Palermo, Italy
e-mail: stefano.decantis@unipa.it; mauro.ferrante@unipa.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_32,

Introduction

Having more and reliable statistics is essential for policy-makers to make effective decisions, for designing marketing strategies, evaluating the efficiency and effectiveness of management decisions, and measuring tourism throughout the regional/local economy. In the last decades there was a growing awareness that the weakness of the statistical data in tourism needed some major initiatives. However, despite the efforts demonstrated by national and international institutions (WTO 1994; European Communities 1994; UNWTO 2010; European Parliament 2011), for improving the reliability and the comparability of statistical information on tourism, current statistics produced by national institutes seem to be still inadequate for destination management purposes, mainly at a local (subregional) level. To date, to answer satisfactory to an apparent simple question such as "how many tourists visited a certain destination each year?" is still an open issue, both under the theoretical and the applied perspective. In a nutshell, simply counting the number of tourists in a destination is not as simple as one might initially think (Smith 1995, p. 16).

At the European level, the partial inadequacy of tourism statistics is demonstrated by the recent new Regulation (EU) No. 692/2011 of the European Parliament and of the Council, concerning European statistics on tourism, which tries to establish a common framework for the systematic development, production and dissemination of European statistics on tourism (European Parliament 2011: art. 1). The new Regulation repeals the Council Directive 95/57/EC and, on the one hand, highlights that the Union's tourism industry occupies an important place in the economy of the Member States, with tourist activities representing a large potential source of employment. On the other hand, the Regulation affirms that any appraisal of its competitiveness requires a good knowledge of the volume of tourism, its characteristics, the profile of the tourist and tourism expenditure and the benefits for the economies of the Member States. It appears that due to: (a) the growing importance of short trips and same-day visits contributing substantially in many regions or countries to the income from tourism, (b) the increasing importance of non-rented accommodation or accommodation in smaller establishments, and (c) the growing impact of the Internet on the booking behaviour of tourists and on the tourism industry, the production of tourism statistics should be adapted and the recommendation 95/57 CE overcame. However, the weakness of tourism statistics highlighted by the European Parliament Regulation, and by several other authors (Lickorish 1997; Vaccina et al. 2011), are not only due to the partial inadequacy of methodologies for the collection of information of the different Member States, but rather the complex nature of tourism phenomenon.

The aim of this paper is to propose a model, named the Trips–Arrivals (T–A) model, for the investigation of the number of tourism trips in a given destination starting from the information on guests arrivals derived from supply-side statistics on collective establishments. After presenting the three main equations (i.e. Night–Presences; Trips–Arrivals; Average Duration of Visit–Average Length of Stay), which jointly constitute the T–A model, some empirical evidences derived from

a survey on incoming tourism in Sicily are presented, demonstrating the inadequacy of official tourism statistics, for planning and management purposes, at regional and subregional level. The T–A model offers a useful framework for destination managers and academics, for the analysis of tourism at the destination level, and for constructing and interpreting adequately tourism indicators, by correcting official accommodation statistics, on overnight stays, arrivals and average length of stay.

The Trips–Arrivals (T–A) Model for the Estimation of Tourism Trips at Local Level

At the European level, information on tourism are collected both from the demand side, and from the supply-side of tourism market. However, statistical sources on the demand side are generally based on sampling surveys that are not designed to give local information (Wanhill 1991, p. 80). This implies that the only available local information are provided from the supply-side statistics on guests in collective establishments. However, they are affected by several problems: first, no information on the motivation of the stay is gathered from the supply-side, making it impossible to distinguish tourists from other guests (e.g. seasonal workers, students, etc., UNWTO 2010, p. 19). Second, not all tourists stay at collective accommodations, and those who do not might have very different patterns of behaviour than those who do. Some kind of accommodations (e.g. non-collective and private accommodations), in fact, are not included in the statistics from the supply-side at all, such as second houses, vacation houses, boats, relatives and friends houses, and so on (Hall and Müller 2004; Gallent and Tewdwr-Jones 2000). We will call this component of tourism demand "unmeasured tourism" (Parroco and Vaccina 2004), according to the terminology used in the field of unobserved economy (OECD 2002). Third, as for many other economic activities, accommodation managers may choose to declare only part of their guests in order to avoid direct or indirect taxation. We will call this component "underground tourism" (Parroco and Vaccina 2004). Fourth, travellers while on a trip might stay in more than one collective accommodation, resulting in an overestimation (i.e. the "double counting effect") of the number of travels and in an underestimation of the total duration of the visit in the destination considered (Pearce 1995; Lickorish 1997; Lickorish and Jenkins 1997; Parroco and Vaccina 2004). Given these considerations, we formalize a conceptual model of actual tourism in a destination/region, by expressing the above-mentioned problems in terms of parameters and/or quantities to be estimated.

The Nights–Presences (N–P) Equation

Let ${}^{\text{obs}}P_{i,t}$ be the total number of nights spent by guests in official collective establishments in the i-destination/region, during the time interval t (e.g. 1 year). Some of these nights can be spent by tourists (in a proportion equal to α_0), and

some other by other kind of guests (e.g. seasonal workers, crews on public modes of transport, students, etc.), according to the UNWTO definition of classification of inbound travellers (UNWTO 2010, p. 17). It follows that the nights spent by tourists in official establishments (${}^{\text{obs}}N_{i,t}$) will be equal to $\alpha_0 {}^{\text{obs}}P_{i,t}$, where $0 \leq \alpha_0 \leq 1$. As stated above, the unobserved tourism may be due both to the nights spent by tourists in unofficial establishments (e.g. private houses, boats, second houses, etc.), and to the nights concealed from public authorities mainly for fiscal reason (OECD 2002, p. 13). We will call the former component "unmeasured tourism" (${}^{\text{unm}}N_{i,t}$), and the latter "underground tourism" (${}^{\text{und}}N_{i,t}$) (Parroco and Vaccina 2004). These two components constitute what we call unobserved tourism (${}^{\text{unobs}}N_{i,t}$). Subsequently, the actual number of nights spent by tourists in a given destination/region i, in the time interval t considered, would be equal to:

$$ {}^{\text{tot}}N_{i,t} = {}^{\text{obs}}N_{i,t} + {}^{\text{unobs}}N_{i,t} = \alpha_0 {}^{\text{obs}}P_{i,t} + {}^{\text{unm}}N_{i,t} + {}^{\text{und}}N_{i,t}, \quad (1) $$

where ${}^{\text{unobs}}N_{i,t}$ represents the unobserved nights spent by tourists in the destination/region i, during the time interval t. For simplicity, we call this expression as the Nights–Presences (N–P) equation. However, if we consider the available official information provided by supply-side statistics, only the first aggregate is known ${}^{\text{obs}}P_{i,t}$. The motivation coefficient (α_0), the unmeasured, and the underground components are unknown, and need to be estimated (for a brief review of the methods proposed for the estimation of these components, see Vaccina et al. 2011).

The Trips–Arrivals (T–A) Equation

Despite the importance of knowing the number of nights spent by tourists in a destination/region, for many planning and management issues, it would be essential to estimate the number of tourism trips made, in a given time interval, in the destination/region considered. The problem of converting available information on guests arrivals into trips is related both with the issues above discussed, but also with the implications of tourist mobility. Subsequently, to convert guests into trips, it is necessary to introduce another coefficient to take into account for the average number of establishments (official and unofficial) used by tourists during their visits in the destination/region considered.

Let ${}^{\text{obs}}G_{i,t} = \alpha_1({}^{\text{obs}}A_{i,t})$ be the number of tourists arrivals registered in official accommodation establishments (where $0 \leq \alpha_1 \leq 1$ represents the proportion of guests arrivals ${}^{\text{obs}}A_{i,t}$ with tourist motivations), in the i-destination/region, during the time interval t; and let β be the average number of establishments used by tourists during their visit within the destination/region considered ($\beta \geq 1$). The number of tourism trips ($\text{TRIPS}_{i,t}$) in the destination/region i, during the time interval t, would be equal to:

$$ \text{TRIPS}_{i,t} = \left({}^{\text{obs}}G_{i,t} + {}^{\text{unobs}}G_{i,t}\right)/\beta = \left[\alpha_1 \left({}^{\text{obs}}A_{i,t}\right) + {}^{\text{unm}}G_{i,t} + {}^{\text{und}}G_{i,t}\right]/\beta, \quad (2) $$

where ${}^{\text{unm}}G_{i,t}$ is the number of tourists which used establishments for which information on arrivals and nights spent are not collected ("unmeasured tourism"); and ${}^{\text{und}}G_{i,t}$ is the number of tourists which used official accommodation establishments, but were not declared to public authorities, mainly for fiscal reasons ("underground tourism"). For simplicity, we call this expression as the Trips–Arrivals (T–A) equation. Supply-side statistics usually provide information only on the number of guests arrivals in official establishments (i.e. ${}^{\text{obs}}A_{i,t}$); in contrast, the remaining aggregates need to be estimated. Moreover, the knowledge of the value of the β coefficient (i.e. the average number of accommodation establishments used by tourists during their stay) becomes really relevant. This issue falls into the broader phenomenon of tourist mobility (Lue et al. 1993; McKercher and Lew 2004); a topic almost ignored by official statistics, but which has important implications not only for the estimation of tourism trips at subregional level, but also for the provisioning and the management of tourism services and logistics.

The Average Duration of Visit–Average Length of Stay (ADOV–ALOS) Equation

Finally, another important aggregate generally considered as an indicator of tourist behaviours is given by the so-called "average length of stay" (ALOS), defined by the ratio between overnight stays and arrivals: ${}^{\text{obs}}\text{ALOS}_{i,t} = {}^{\text{obs}}P_{i,t}/{}^{\text{obs}}A_{i,t}$. This index is usually used as an indicator of the length of the trips in the destination/region considered. However, for the problems above highlighted (i.e. unobserved tourism and tourists mobility), it is only a measure of the "average length of stay" in official accommodation establishments. In contrast, the "average duration of visit" (ADOV), in the destination/region i, during the time interval t considered, according to our framework would be given by the ratio between Eqs. (1) and (2). Subsequently, we have:

$$\text{ADOV}_{i,t} = \frac{{}^{\text{tot}}N_{i,t}}{\text{TRIPS}_{i,t}} = \beta \frac{\alpha_0 \left({}^{\text{obs}}P_{i,t}\right) + {}^{\text{unm}}N_{i,t} + {}^{\text{und}}N_{i,t}}{\alpha_1 \left({}^{\text{obs}}A_{i,t}\right) + {}^{\text{unm}}G_{i,t} + {}^{\text{und}}G_{i,t}} \tag{3a}$$

$$= \beta \frac{\left[\frac{\alpha_0}{\alpha_1}{}^{\text{obs}}\text{ALOS}_{i,t} \times \alpha_1 \left({}^{\text{obs}}A_{i,t}\right)\right] + \left({}^{\text{unm}}\text{ALOS}_{i,t} \times {}^{\text{unm}}G_{i,t}\right) + \left({}^{\text{und}}\text{ALOS}_{i,t} \times {}^{\text{und}}G_{i,t}\right)}{\alpha_1 \left({}^{\text{obs}}A_{i,t}\right) + {}^{\text{unm}}G_{i,t} + {}^{\text{und}}G_{i,t}}. \tag{3b}$$

We call this expression the Average Duration of Visit–Average Length of Stay (ADOV–ALOS) equation. It expresses the average duration of visit, in the destination/region i, during the time interval t considered ($\text{ADOV}_{i,t}$), as a weighted mean of the average length of stays (of tourism trips) in the different types of establishments/situations considered (official, unmeasured, underground), multiplied by the β coefficient—where the weights are given by the number of

guests arrivals (official, unmeasured, underground). If the β coefficient would not be taken into account, the weighted mean of the different ALOS represents the average length of stay in the different establishments/situations considered (official, unmeasured, underground). Official accommodation statistics only provide information on ${}^{obs}ALOS_{i,t}$.

Empirical Evidences from the Survey on Incoming Tourism in Sicily

In the period between 2009 and 2010, thanks to a research project co-founded by the Italian Ministry of University and Research, the research group of the University of Palermo and Catania, composed mainly by social statisticians, planned a survey covering the whole Sicily. The survey aimed at estimating the actual magnitude of tourism in the island, by quantifying two of the main biases of the statistics on guests arrivals: the double counting effect (i.e. the β parameter), and the unobserved tourism (particularly, the unmeasured component, i.e. ${}^{unm}G_{i,t}$, and ${}^{unm}N_{i,t}$). The survey was designed only for incoming tourists in Sicily, since they represent the most important segment of tourism market in the island. Moreover, a survey on domestic tourism would require a different sampling design strategy and specific interview techniques (e.g. Telephone surveys). The survey used a complex Time Location Sampling (TLS) design, given the mobile and the special nature of tourists population (see De Cantis et al. 2010; Kalsbeek 2003). The units of interest were represented by Italian (not resident in the island) and foreign tourists leaving Sicily at the end of their vacation. Thus, it was possible to collect direct information (from the demand side) related to the whole period spent in Sicily, through a direct interview, allowing to reduce the recall bias, which usually affects many demand-side surveys (Rylander et al. 1995). A detailed description of the sampling design is contained in De Cantis et al. (2010). The insularity of Sicily allowed us to select almost all the places from which it is possible to leave the island, namely: the airports of Palermo, Catania, and Trapani, the ports of Palermo and Catania, and the Strait of Messina (only the two airports of the two small islands Pantelleria and Lampedusa were not included in the survey). The periods covered by the survey (i.e. Spring, Summer, and Autumn) were selected according to official data on tourists flows in the island, covering more than the 80 % of official incoming tourists flows. The research group designed specifically a questionnaire of 29 items, divided into different sections: filter questions and organization of the trip; motivations and expectations; type of holyday (sea and sand, cultural, etc.); intra-regional mobility and type of establishments used; expenses; satisfaction. The specific section related to the collection of information on tourism mobility and on unobserved tourism is presented in Fig. 1.

In this section, the tourist was asked to specify all the places (municipalities) which he/she visited during his/her trip, with at least one overnight stay. For each stay, he/she was asked to specify the number of nights spent, and the types

Places visited	Nights spent	Official establishments						Unofficial establishments			
		01. Rural facilities	02. Holyday or work camp	03. Hotels and similar establishments	04. Camping	05. Bed & Breakfast	06. Youth Hostel	07. House/Room rented	08. Relatives or friends house	09. Second home	10. Other (specify)
		☐	☐	☐	☐	☐	☐	☐	☐	☐	…
		☐	☐	☐	☐	☐	☐	☐	☐	☐	…
		☐	☐	☐	☐	☐	☐	☐	☐	☐	…

Fig. 1 Questionnaire section on tourism mobility

of accommodation establishments used, in order to distinguish between official and unofficial establishments. Through this section it was possible to relate the information collected with the two topics of interest: tourist mobility (i.e. the β parameter) and unmeasured tourism (i.e. ${}^{\text{unm}}G_{i,t}$, and ${}^{\text{unm}}N_{i,t}$).

Between Summer 2009 and Spring 2010 a total of 3,935 valid cases were collected (i.e. incoming tourists in Sicily). The complex sample design is described in De Cantis, et al. (2010). The sample was constituted for about 60 % by Italian tourists and the remaining 40 % by foreigners. Regarding the type of vacation, only 21.5 % made a sea and sand holiday at all, for the 52.7 % of tourists, their holiday was only partially sea and sand, whereas for remaining 25.6 % made other types of vacation. By remanding to a forthcoming work (Oliveri and De Cantis 2013) the presentation of the final results which take into account for the sampling design adopted, in this paper we present raw sample data in order to implement the T–A model, by highlighting both the unmeasured component of tourism demand in Sicily, and the effects produced by tourist mobility in the island. In Table 1, the raw results related with tourist mobility and with on unobserved tourism are reported.

According to the T–A model, it is possible to quantify some of the aggregates contained in the three equations of the model. Regarding the N–P equation, a total of 38,644 nights (${}^{\text{tot}}N_{i,t}$) were spent by the incoming tourists sampled in Sicily, during the time interval considered. These are only partially measured by official statistics on guest arrivals, since unofficial establishments are not covered by supply-side statistics (${}^{\text{unm}}N_{i,t} = 22{,}113$). Thus, the share of unmeasured overnight stays is about 57 % of the total nights spent in Sicily by the interviewed tourists. Moreover, the nights spent in official establishments may be partially concealed to public authorities for fiscal reasons, despite the approach undertaken does not allow to separate the underground component from the official one, turning out in a total of 16,531 nights ($\alpha_0 {}^{\text{obs}}P_{i,t} + {}^{\text{und}}N_{i,t} = 16{,}531$). By considering the second equation (i.e. the T–A equation) of the model, a total of 3,935 tourism trips ($\text{TRIPS}_{i,t}$) sampled produced 6,485 stays, distributed among several establishments categories, which determines a value of the average number of stays (β) almost equal to 1.65 (we assumed that tourists do not change accommodation establishment within a single municipality). The stays in official establishments (i.e. $\alpha_1 {}^{\text{obs}}A_{i,t} + {}^{\text{und}}G_{i,t}$) were equal

Table 1 Results in terms of stays and overnight stays by accommodation establishment category, from 3,935 interviews to incoming tourists in Sicily, Summer–Autumn 2009, Spring 2010

Accommodation establishment category		Stays	Overnight stays	Average length of stay
Official establishments	Rural establishments	152	589	3.88
	Holyday camps	24	200	8.33
	Hotels	2,615	11,071	4.23
	Camping	377	1,183	3.14
	Bed and breakfast	1,023	3,359	3.28
	Youth hostels	46	129	2.80
Unofficial establishments	House or room rented	461	4,607	9.99
	Relative and friends houses	1,354	12,587	9.30
	Owned houses	307	4,502	14.66
	Other unofficial establishments	126	417	3.31
Total		6,485	38,644	5.96
Average duration of visit in Sicily ($ADOV_{i,t}$) = 38,644/3,935 = 9.82				

Table 2 Arrivals, overnight stays, and average length of stay of non-Sicilian guests (residents in other Italian regions and Foreigners) in official accommodation establishments in Sicily, year 2009

Establishment category	Arrivals	Overnight stays	Average length of stay
Hotels and similar establishments	2,491,373	8,325,020	3.34
Other collective establishments	319,508	1,320,144	4.13
Total	2,810,881	9,645,164	3.43

Source: Osservatorio Turistico Regione Siciliana (2011)

to 4,237, whereas the unmeasured stays ($^{unm}G_{i,t}$) were 2,248. Finally, regarding the ADOV–ALOS equation, the average duration of visit ($ADOV_{i,t}$) in Sicily is equal to 9.82 nights; the average length of stay ($ALOS_{i,t}$) varies strongly among the different establishment categories, being 3.90 nights and 11.65 nights, in official and unofficial establishments, respectively.

As above highlighted, official data on guests in accommodation establishments provide information only on arrivals, presences, and average length of stay in official establishments, as shown in Table 2. The average length of stay obtained from official data (Table 2) is in line with the sampling results related with the average length of stay in official establishments, derived from the survey on incoming tourists (Table 1). This example empirically shows the biases which occur when data on guests arrivals, overnight stays and average length of stay, are used to quantify and characterize tourism flows in a given destination.

Comments and Conclusions

Increasingly regional tourism authorities are interested in regional statistics. However, as highlighted in this work, at subregional and local level, demand-side statistics are not provided by the European system of tourism statistics. This determined the habits of destination managers, and researchers, of using accommodation statistics to analyse tourism demand. However, the use of supply-side information to examine demand-side features may determine conceptual and practical mistakes. Some of these issues are getting recognized by major institutions, such as the European Travel Commission (ETC) and Eurostat. For example, the recent quarterly report published by the ETC on European tourism-trend and prospects (ETC 2010), in comparing the results of the US Department of Commerce, reporting a decline of US outbound travel to Europe, with the results of the TourMIS, which indicated an increase of US arrivals in Europe and a reduction of the average length of stay, commented that: "one plausible way to read the data is that US travellers are participating in multi-leg European trips with shorter stay in each destination" (ETC 2010, p. 16). However, despite the phenomenon of multi-destination trip is being recognized, there are still no official sources of information which allow to measure its magnitude and its features, neither at the national level (visits to several regions, municipalities, etc.), nor at the international one (visits to several countries).

Concerning the unobserved component of tourism, related to the use of unofficial establishments (namely, the unmeasured tourism), the new Regulation (EU) No. 692/2011 of the European Parliament puts emphasis on the non-rented accommodations, meaning, inter alia, accommodations provided without charge by family or friends and accommodations in owner-occupied vacation homes, including time share properties (European Parliament 2011, p. 19). According to the new Regulation, the data to be transmitted by the Member States shall relate not only to the capacity and occupancy of tourists accommodation establishments, but also to tourism nights spent in non-rented accommodation. To date, however, nor in Italy, nor in the other European countries, information on unmeasured tourism (and on underground tourism) are available, and the way in which member countries will collect and provide these information is still an open issue.

The T–A model allows to face with the problem of quantifying the number of tourism trips made in a given destination by highlighting the lack of official supply-side tourism statistics. However, this approach presents some limits which are mainly related with the estimation of the parameters and quantities presented in the equations. With reference to the motivation coefficients (i.e. $\alpha_{0,}$ and α_1) it should be kept in mind the characterization of the destination/region. In tourism resorts, it could be assumed that all guests are tourists (i.e. $\alpha_0 = \alpha_1 = 1$); however, this hypothesis would be unreliable in urban destinations where other guests (e.g. workers) are likely to visit the destination and stay in collective establishments. In these cases, an estimate of α_0 and α_1, obtained, for example, through a sample survey in official establishments, would be required. The unmeasured component of tourism demand is closely related to the presence of the so-called "unofficial

establishments", such as second houses, rooms or houses rented. The estimation of the number of second houses in a given destination/region, for example through information coming from the census on population and housing, could help to understand the potential magnitude of the unmeasured tourism in the destination considered. The underground component is even harder to quantify, since deliberately concealed to public authorities. This issue falls into the broader issue of measuring unobserved economy (OECD 2002), and, to date, no direct solutions have been proposed. Finally, regarding the β parameter, next to nothing is known about the number of destinations visited by tourists. However, for small areas, such as municipalities a value of $\beta = 1$ could be assumed, whereas for larger areas, such as tourism districts, or Provinces, an estimate of β would be required. By concluding, a deeper knowledge of tourists behaviour is required to determine the values of these parameters and the factors affecting their variability in time and space. Furthermore, the changing nature of demand and the increasing segmentation of the holyday market are also raising the need for more accurate, destination-based, information which integrate quantitative information on the magnitude of tourism with other more specific aspects of tourism behaviours.

References

De Cantis, S., Gonano, G., Scalone, F., Vaccina, F.: Il disegno campionario e il piano di rilevazione nell'indagine sui turisti incoming in partenza dalla Sicilia e dalla Sardegna: il campionamento spazio-temporale per popolazioni hard to reach. In: Parroco, A.M., Vaccina, F. (eds.) Mobilità ed altri comportamenti dei turisti: studi e ricerche a confronto, pp. 21–46. McGraw-Hill, Milano (2010)

European Communities: Community methodology on tourism statistics, Office for Official Publications of the European Communities, Luxembourg (1994)

European Parliament: Regulation (EU) No. 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning European statistics on tourism and repealing Council Directive 95/57/EC. Official Journal of the European Union, L192 (54), pp. 17–32 (2011)

European Travel Commission: European Tourism in 2010: Trends and Prospects, Quarterly Report Q3/2010, European Travel Commission, Brussels (2010)

Gallent, N., Tewdwr-Jones, M.: Rural Second Homes in Europe: Examining Housing Supply and Planning Control. Ashgate, Aldershot (2000)

Hall, C.M., Dieter, K.: Tourism, Mobility, and Second Homes: Between Elite Landscape and Common Ground. Channel View Publications, Buffalo (2004)

Kalsbeek, W.D.: Sampling minority groups in health surveys. Stat. Med. **22**, 1527–1549 (2003)

Lickorish, L.J.: Travel statistics – the slow move forward. Tourism Manag. **18**(8), 491–497 (1997)

Lickorish, L.J., Jenkins, C.L.: An Introduction to Tourism. Butterworth–Heinemann, Oxford (1997)

Lue, C.C., Crompton, J.L., Fesenmaier, D.R.: Conceptualization of multi-destination pleasure trips. Ann. Tourism Res. **20**, 289–301 (1993)

McKercher, B., Lew, A.A.: Tourist flows and the spatial distribution of tourists. In: Lew, A.A., Hall, M., Williams, A.M. (eds.) A Companion to Tourism, pp. 36–48. Blackwell Publishing, Malden (2004)

OECD: Measuring the Non-observed Economy. A Handbook. OECD Publication Services, Paris (2002)

Oliveri, A.M., De Cantis, S.: Mobilità del turismo regionale incoming. Aspetti socio-economici dei comportamenti e delle motivazioni. Collana di Scienze del Turismo, vol.3, McGraw-Hill, Milano (2013)

Parroco, A.M., Vaccina, F.: Estimates of hidden tourism to plan local services: the Sicilian case. In: SCORUS Conference Proceedings, Scorus, Minneapolis, (2004)

Pearce, D.: Tourism Today. A Geographical Analysis, 2nd edn. Longman, Harlow (1995)

Osservatorio Turistico della Regione Siciliana: Movimenti turistici nella Regione, anno 2009. Database available on http://pti.regione.sicilia.it/portal/page/portal/PIR_PORTALE/PIR_LaStrutturaRegionale/PIR_TurismoSportSpettacolo/PIR_Turismo/PIR_7338501.618136477/flussi_cat_2009-10.xls (2011). Last access on 18 Sept 2012

Rylander II, R.G., Propst, D.B., McMurtry, T.R.: Non-response and recall biases in a survey of traveler spending. J. Travel Re. **33**(4), 39–45 (1995)

Smith, S.L.: Tourism Analysis: A Handbook. Longman Group, Essex (1995)

UNWTO: International Recommendations for Tourism Statistics 2008, United Nations Publications, Studies in Methods, Series M, 83(1), New York. Available on http://unstats.un.org/unsd/publication/Seriesm/SeriesM_83rev1e.pdf (2010). Last access on 18 Sept 2012

Vaccina, F., Parroco, A.M., De Cantis, S., Ferrante, M.: Un-observed tourism: approaches and case studies in Sicily. In: Proceedings of the TTRA Europe 2011 and AFM Conference "Creativity and innovation in tourism", Technopole d'Archamps, 11–13 April (2011)

Wanhill, S.: UK Visitor Survey, Tourism Management, March, pp. 79–80 (1991)

WTO: Recommendations on Tourism Statistics. WTO, Madrid (1994)

Non-compensatory Aggregation of Social Indicators: An Icon Representation

Matteo Mazziotta and Adriano Pareto

Abstract

In this paper, we consider a non-compensatory composite index, denoted as MPI (Mazziotta–Pareto Index) and propose an original method for visualizing the index value for a set of statistical units. The MPI is characterized by two elements: "mean" and "penalty". The idea is to represent each unit as a particular graphical object, a "stickman with a sack", where the value of the "mean" is assigned to the size of the "stickman" and the value of the "penalty" is assigned to the size of the "sack". The assignment is such that the overall appearance of the object changes as a function of the MPI values.

Introduction

Composite indices for comparing country performance with respect to multi-dimensional phenomena, such as socio-economic development, poverty, quality of life, etc., are increasingly recognized as a useful tool in policy and public communication (OECD 2008).

Considerable attention has been devoted in recent years to the fundamental issue of compensability among the components of the index (a deficit in one dimension can be compensated by a surplus in another) and more and more often a non-compensatory approach has been adopted. For example, the "new" Human Development Index calculated by the United Nations Development Programme in 2010 is given by a geometric mean (UNDP 2010).

M. Mazziotta • A. Pareto (✉)
Italian National Institute of Statistics, Rome, Italy
e-mail: pareto@istat.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_33,

The aim of this work is to provide an original graphical method, called "Traveller Icon" plot, for representing the non-compensatory aggregation of individual indicators by MPI (De Muro et al. 2010). The proposed method allows not only visualizing and analyzing the message contained in data, but also remembering it, since for most of people, visual memory is more persistent than verbal or auditory memory (Zinovyev 2011). For that reason, it can serve as a powerful propagandistic or educational tool.

In Sect. 2, a brief description of MPI is reported; in Sect. 3, the "Traveller Icon" plots are presented; finally, in Sect. 4, an application to real data is proposed.

A Non-compensatory Composite Index

The MPI (Mazziotta–Pareto Index) is a composite index based on the assumption of "non-substitutability" of the indicators, i.e., they have all the same importance and a compensation among them is not allowed (De Muro et al. 2010).

The steps for computing MPI are given below.

Given the matrix $\mathbf{X} = \{x_{ij}\}$ with n rows (units) and m columns (indicators), we calculate the standardized matrix $\mathbf{Z} = \{z_{ij}\}$ as follow:

$$z_{ij} = 100 \pm \frac{(x_{ij} - M_{x_j})}{S_{x_j}} 10$$

where M_{x_j} and S_{x_j} are, respectively, the mean and the standard deviation of the jth indicator and the sign $\pm$ is the "polarity" of the jth indicator, i.e., the sign of the relation between the jth indicator and the phenomenon to be measured ($+$ if the individual indicator represents a dimension considered positive and $-$ if it represents a dimension considered negative).

Denoting with M_{z_i} and S_{z_i}, respectively, the mean and the standard deviation of the standardized values of the ith unit, the generalized form[1] of MPI is given by:

$$\text{MPI}_i^{+/-} = M_{z_i} \pm S_{z_i} cv_i$$

where $cv_i = S_{z_i}/M_{z_i}$ is the coefficient of variation of the ith unit and the sign $\pm$ depends on the kind of phenomenon to be measured.

If the composite index is "increasing" or "positive", i.e., increasing values of the index correspond to positive variations of the phenomenon (e.g., the socio-economic development), then MPI^- is used. Vice versa, if the composite index is "decreasing" or "negative", i.e., increasing values of the index correspond to negative variations

[1] It is a generalized form since it includes "two indices in one".

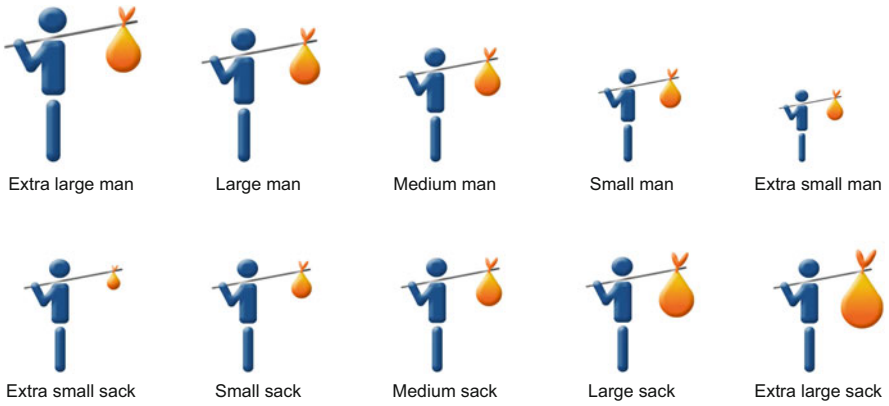

Fig. 1 Examples of "Traveller Icon"

of the phenomenon (e.g., the poverty), then MPI$^+$ is used. For some applications, see Mazziotta and Pareto (2010), De Muro et al. (2010).

This approach is characterized by the use of a function (the product $S_{z_i} cv_i$) to penalize the units with "unbalanced" values of the indicators. The "penalty" is based on the coefficient of variation and is zero if all the values are equal.

The purpose is to reward the units that, mean being equal, have a greater balance among the different indicators. Therefore, the MPI is characterized by the combination of a "mean effect" (M_{z_i}) and a "penalty effect" ($S_{z_i} cv_i$).

The "Traveller Icon" Plots

The basic idea of "Traveller Icon" plots is to represent each unit as a "stickman with a sack", where the value of the "mean" of the standardized indicators is assigned to the size of the "stickman" and the value of the "penalty" is assigned to the size of the "sack". In this way, the overall appearance of the icon changes as a function of the MPI values.

Figure 1 shows some examples of "Traveller Icon" for decreasing values of the "mean" ("stickman" size) and increasing values of the "penalty" ("sack" size). All the combinations of "sacks" and "stickmen" are possible.

Examining such icons may help to discover interactions between "mean effect" and "penalty effect" and identify specific clusters of units (e.g. units with high values of "penalty" are represented by "stickmen with a large sack", whereas units with low values are represented by "stickmen with a small sack").

An Example of Application

In order to show the graphical representation of the MPI by "Traveller Icon" plots, an application is presented where a set of indicators of Quality of Life (QoL) in the Italian cities, at regional level, is considered.

The variables used are the following: Sporting activities, Distance to supermarkets, Air quality, Urban crime, Green space, Public transport, Parking provision, Attractiveness of universities, Attractiveness of health services, Children's services. Each indicator is interpreted as "positive" or "negative" with respect to QoL (polarity).

The MPI^{-} is used, since the composite index is "positive", i.e., increasing values of the index correspond to positive variations of QoL.

The data matrix is shown in Table 1, where the original values and the standardized values of the individual indicators are reported (the polarities are between brackets).

Table 2 reports the ranking of the Italian regions by MPI^{-}. Also provided in the table are the mean of the standardized values of indicators and the penalty.

On the basis of the distributions of values across regions, five frequency classes for the mean (<95; 95–100; 100–102; 102–104; >104) and the penalty (<0.25; 0.25–0.5; 0.5–1; 1–2.5; >2.5) were fixed. The bar graphs of the frequencies are displayed in Fig. 2. Note that the limits of the classes for the "mean effect" are very close to the quintiles.

The traveller plots are displayed in Fig. 3, where larger sizes of the "stickman" correspond to increasing classes of the "mean" and larger sizes of the "sack" correspond to increasing classes of "penalty".[2]

It is observed from the figure that Valle d'Aosta (rank 11) has a high value of the mean (extra large man), but a high value of the penalty (extra large sack) too; indeed its ranking is much lower compared to Veneto (rank 3).

Umbria and Marche are rather similar regions in terms of both mean and penalty. Finally, Molise (rank 20) and Basilicata (rank 19) have low values of the mean (extra small man) and high values of penalty (large sack).

Conclusions

The main objective of data visualization is to provide an efficient graphical display for summarizing quantitative information.

Data visualization plays a very important role in social sciences, since it helps to create informative illustrations of the data, representing large amount of quantitative information in a form that is much more accessible to the human eye.

Multivariate plots, specifically, are designed to visualize the values of several variables at the same time and allow comparison among different units. Examples of

[2]Note that the higher the number of classes, the greater the accuracy of the icons.

Table 1 Individual indicators of QoL in the Italian cities—years 2003–2008[a]

Region	Sporting activities (+)	Distance to supermarkets (−)	Air quality (+)	Urban crime (−)	Green space (+)	Public transport (+)	Parking provision (+)	Attractiveness of universities (+)	Attractiveness of health services (−)	Children's services (+)
Original values										
Piemonte	34.1	31.0	1.9	29.5	42.5	199.3	17.1	−7.8	5.7	37.1
Valle d'Aosta	46.3	41.4	10.5	12.1	26.2	580.0	8.4	−214.8	14.4	78.4
Lombardia	36.5	31.1	1.8	30.6	28.6	227.7	24.1	8.3	3.8	62.5
Trentino-Alto Adige	48.2	28.1	2.0	12.7	70.3	192.9	34.5	−22.0	9.4	83.8
Veneto	39.6	29.9	1.4	19.7	62.3	124.4	42.2	−8.7	3.3	70.2
Friuli-Venezia Giulia	37.5	25.4	3.2	8.4	22.1	258.1	12.0	8.3	5.6	83.6
Liguria	27.6	29.4	4.2	17.2	35.4	311.0	22.3	−9.2	8.5	64.3
Emilia-Romagna	36.8	30.7	2.5	21.4	157.7	83.0	24.0	33.7	5.0	88.0
Toscana	33.1	35.7	1.9	16.2	152.1	108.4	20.9	18.0	4.5	74.6
Umbria	32.3	26.3	1.9	10.6	187.6	162.8	26.9	23.6	11.5	63.0
Marche	32.2	32.6	2.3	5.9	186.1	157.7	15.3	0.3	8.2	55.7
Lazio	29.4	25.3	1.1	30.4	121.0	132.3	7.0	24.3	4.6	30.7
Abruzzo	31.0	37.0	0.7	8.2	710.0	93.5	21.1	29.4	10.5	52.1
Molise	22.0	41.3	0.0	4.8	18.5	177.2	1.2	−41.4	18.7	7.4
Campania	21.1	40.0	0.4	23.2	25.9	218.0	5.9	−17.9	9.9	50.5
Puglia	23.8	30.4	1.4	14.3	8.1	122.0	8.2	−38.9	7.9	44.2
Basilicata	27.1	34.8	1.2	4.4	545.6	87.4	2.3	−201.1	22.2	21.4
Calabria	24.8	43.6	0.3	15.1	20.8	172.8	19.5	−57.2	16.2	15.6
Sicilia	22.5	31.4	1.3	14.4	73.3	75.7	6.5	−10.6	7.5	34.6
Sardegna	28.2	21.7	3.6	14.3	85.9	56.6	16.9	−23.7	4.7	20.4

(Continued)

Table 1 (Continued)

Region	Sporting activities (+)	Distance to supermarkets (−)	Air quality (+)	Urban crime (−)	Green space (+)	Public transport (+)	Parking provision (+)	Attractiveness of universities (+)	Attractiveness of health services (−)	Children's services (+)
Standardized values										
Piemonte	103.2	102.3	98.7	82.5	95.1	102.0	100.3	102.7	106.7	93.8
Valle d'Aosta	119.8	84.5	138.3	104.5	94.2	135.8	91.9	71.0	89.7	111.0
Lombardia	106.5	102.2	98.4	81.0	94.3	104.5	107.0	105.2	110.4	104.4
Trentino-Alto Adige	122.3	107.4	99.4	103.8	96.7	101.4	116.9	100.5	99.4	113.3
Veneto	110.7	104.2	96.5	94.9	96.2	95.3	124.3	102.5	111.4	107.6
Friuli-Venezia Giulia	107.8	111.9	104.8	109.2	94.0	107.2	95.4	105.1	106.8	113.2
Liguria	94.5	105.1	109.2	98.1	94.7	111.9	105.3	102.5	101.2	105.1
Emilia-Romagna	106.9	102.9	101.4	92.7	101.6	91.6	106.8	109.0	108.1	115.0
Toscana	101.8	94.2	98.9	99.3	101.3	93.9	103.9	106.6	109.0	109.4
Umbria	100.8	110.4	98.5	106.4	103.3	98.7	109.6	107.5	95.3	104.6
Marche	100.6	99.5	100.5	112.5	103.2	98.3	98.5	103.9	101.8	101.6
Lazio	96.9	112.1	94.9	81.3	99.5	96.0	90.6	107.6	108.8	91.2
Abruzzo	99.0	92.1	93.2	109.5	132.7	92.6	104.1	108.4	97.3	100.1
Molise	86.9	84.7	90.0	113.8	93.8	100.0	85.1	97.5	81.1	81.4
Campania	85.7	86.9	91.9	90.4	94.2	103.6	89.6	101.1	98.4	99.4
Puglia	89.3	103.4	96.2	101.7	93.2	95.1	91.8	97.9	102.4	96.8
Basilicata	93.7	95.7	95.4	114.3	123.5	92.0	86.1	73.1	74.3	87.3
Calabria	90.7	80.7	91.6	100.8	93.9	99.6	102.6	95.1	86.1	84.9
Sicilia	87.5	101.6	95.9	101.6	96.9	91.0	90.1	102.3	103.2	92.8
Sardegna	95.3	118.3	106.4	101.8	97.6	89.3	100.1	100.3	108.6	86.9

[a]*Data source*: http://www.istat.it/it/files/2011/07/AsseV.xls

Table 2 Italian regions ranking by MPI^{-}

Region	Mean	Penalty	MPI	Rank
Piemonte	98.74	0.43	98.30	13
Valle d'Aosta	104.07	4.23	99.84	11
Lombardia	101.38	0.64	100.74	10
Trentino-Alto Adige	106.10	0.63	105.47	1
Veneto	104.38	0.77	103.61	3
Friuli-Venezia Giulia	105.55	0.34	105.21	2
Liguria	102.76	0.29	102.47	6
Emilia-Romagna	103.62	0.46	103.16	5
Toscana	101.84	0.27	101.57	9
Umbria	103.52	0.22	103.30	4
Marche	102.05	0.15	101.90	7
Lazio	97.88	0.82	97.06	14
Abruzzo	102.90	1.30	101.60	8
Molise	91.43	1.02	90.42	20
Campania	94.12	0.37	93.75	17
Puglia	96.78	0.21	96.58	15
Basilicata	93.55	2.37	91.18	19
Calabria	92.59	0.51	92.08	18
Sicilia	96.29	0.31	95.98	16
Sardegna	100.45	0.76	99.69	12

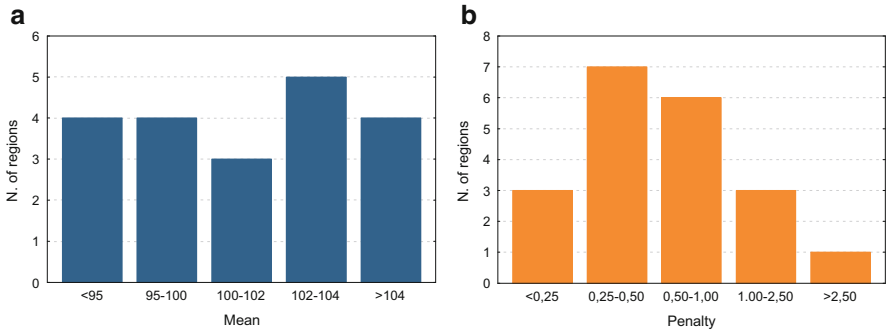

Fig. 2 Bar-diagram for "mean effect" (**a**) and "penalty effect" (**b**)

these are radar charts, sun ray plots and more original Chernoff faces plots (Everitt et al. 1975).

In this work, we present a new and original visual technique, called "Traveller Icon" plot, for representing the non-compensatory aggregation of individual indicators by MPI. The basic idea of "Traveller Icon" plots is to represent each unit as a

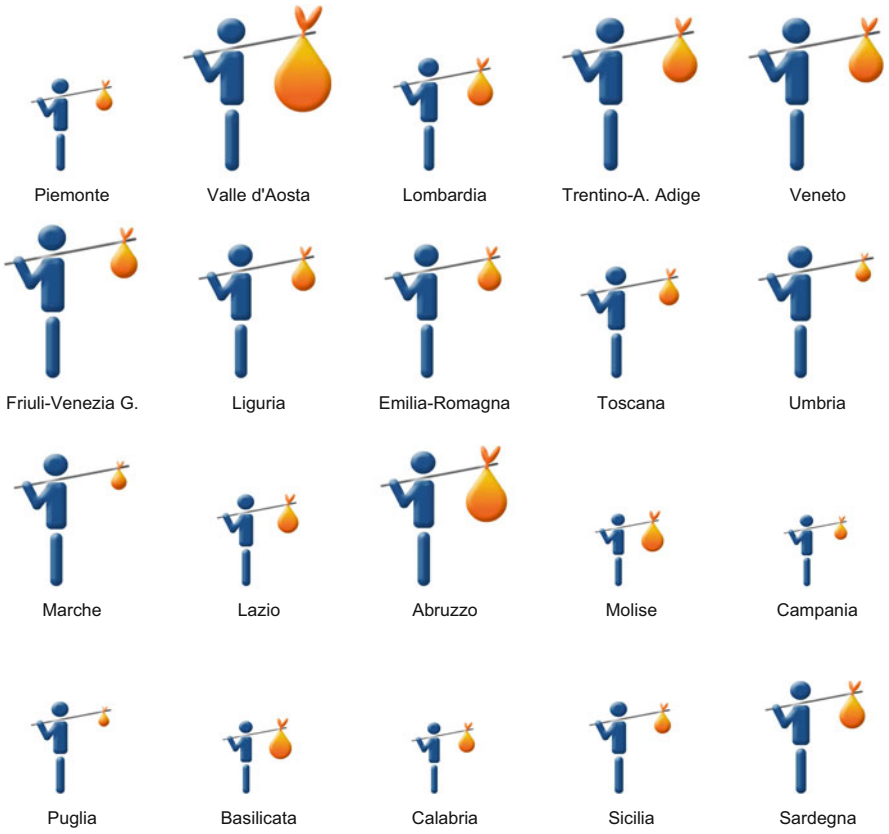

Fig. 3 Traveller plots of the Italian regions by MPI^{-}

"stickman with a sack", where the sizes of the "stickman" and the "sack" depend, respectively, on the values of "mean" and "penalty".

The proposed method may be extended for the representation of any couple of variables.

Acknowledgments Matteo Mazziotta has written Sects. 1 and 2, Adriano Pareto has written Sects. 3–5. We are grateful to Silvia Pareto for designing the graphics.

References

De Muro, P., Mazziotta, M., Pareto, A.: Composite indices of development and poverty: an application to MDGs. Soc. Indic. Res. (2010). doi: 10.1007/s11205-010-9727-z

Everitt, B.S., Nicholls, P.: Visual techniques for representing multivariate data. The Statistician **24**, 37–49 (1975)

Mazziotta, M., Pareto, A.: Measuring quality of life: an approach based on the non-substitutability of indicators. Statistica Applicazioni **8**, 169–180 (2010)

OECD: Handbook on Constructing Composite Indicators. Methodology and User Guide. OECD Publications, Paris (2008)

UNDP: Human Development Report 2010. The Real Wealth of Nations: Pathways to Human Development. Palgrave Macmillan, New York (2010)

Zinovyev, A.: Data visualization in political and social sciences. In: Badie, B., Berg-Schlosser, D., Morlino, L.A. (eds.) International Encyclopedia of Political Science. Sage, London (2011). E-print: arXiv:1008.1188

Indicators for Assessment in Health Services

Cesare Cislaghi and Marco Marchi

Abstract

The transition from a set of healthcare statistics to a Health Information System took place in Italy following the reform of National Statistical Institute (ISTAT) and the establishment of the National Health Service. The introduction of epidemiological monitoring was the characterising element of the new information cycle, with different needs of aetiological type (research on disease determinants) and of evaluation type (efficacy and efficiency). The definition of an indicator set for government of the healthcare system was the object of several laws dealing with public health indicators, cost and expenditure indicators, realisation and appropriateness indicators, performance and outcome indicators. The final section was dedicated to the definition, compilation and use of such indicators.

Statistical-Health Surveying in Italy

When publication of the "Annual Health Statistics" began in 1955, Italian statistical-health surveys were, in many ways, a dis-homogenous and dis-organic set.

Along with certain flows which were also of demographic interest there was also observation deriving from purposes to a certain extent of "health policing" (reports of abortion, reports of births with congenital deformities) with evident, sensational differences in terms of coverage of the phenomena and reliability of the data collected.

C. Cislaghi
Agenzia Nazionale per i Servizi Sanitari Regionali, Rome, Italy

M. Marchi (✉)
Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università di Firenze, Florence, Italy
e-mail: marchi@disia.unifi.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_34,

In the late 1970s, the institution of the National Health Service—NHS (Law No. 833/78, called the Health Reform) imposed new requirements for intervention, participation and research which brought about, the identification and expression, by new actors, of greater and different information requirements.

With the suppression of the precedent mutual aid system, new competencies were indicated for the State (Article 6) , functions assigned to the Regions (Article 7), the structure and function of Local Health Units (LHU) was established (Articles 14–15), the organisation of prevention activities defined (Articles 20–21), but, above all, an "Epidemiological and Statistical Service" was instituted (Article 58) and indeed the National Health Plans (Article 53) were to include "specific programmes for the survey and management of epidemiological, statistical and financial information".

The new health information flows should then have been characterised by a process of "connotation" that is:

(a) Rising from the periphery to the centre as a report of "pertinence" (i.e. report of a problem) then subject to verification of "relevance" (i.e. of quantitative consistency);
(b) Being handed down from the centre to the periphery as an indication of "relevance (i.e. limits exceeded or alarms triggered) to obtain an assessment of "pertinence" (i.e. recognition of importance).

The information exigencies thus created demand for new and/or diverse surveys, with consequent changes to the phases of collection, elaboration and publication of the data, with a view, above all, to their diverse utilisation.

From the Set of Health Statistics to Health Information System

In the opening speech to the ordinary general meeting of the Italian Association of Epidemiology, Rome 16/12/1977 (Marchi et al. 1978) punctual critical observations were raised on both the set up and execution and on the content of the then set of health statistics observations.

(a) Criticism of the SET UP
    1. Exaggerated centralisation;
    2. Bureaucratic hierarchisation;
    3. Separation from users.
(b) Criticism of the EXECUTION
    1. Delays in data processing;
    2. Inadequate dis-aggregations;
    3. Insufficient flexibility.
(c) Criticism of the CONTENT
    1. Administrative or "health police" derivation;
    2. Search for formal accuracy;
    3. Lack of connection between the diverse observations;
    4. Attention to terminal phases of morbid processes;
    5. Prevalent reference to institutionalised situation;
    6. Inadequate research into primary causes;

7. Exclusion of "subjective" components of assessment of the state of health.

The specific points to be pursued under a technical-methodological profile were also re-stated:

UNIQUENESS—to guarantee data homogeneity and comparability;

COORDINATION—to eliminate duplication and deformity of surveying;

RELIABILITY—to reduce the elements of doubt or distortion;

COMPLETENESS—to impede statistical defection and loss of cases;

TIMELINESS—to receive non-obsolete elaborations, adequate, thus, to intervention;

FLEXIBILITY—to enable orientation of the system towards emerging problems;

REFERABILITY—to correctly relate the cases to the population at risk.

Above all it was emphasised that the insertion of the epidemiological observation function, with its' specific characteristics, should have been the key element for the hoped reorganisation for SET of health statistics on the INFORMATION SYSTEM for the new NHS:

FROM	TO
Static	Dynamic
Rigid	Flexible
Centralistic	Decentred
Bureaucratic	Participative
Detached from references to the territory	Adherent to problems in the territory
With individual reference only	Also with reference to the Group
Addressing damages	Addressing risks
Denotative	Connotative
Descriptive	Interpretative
Cognitive (abstract knowledge)	Evaluative (aimed knowledge)
Relevant (quantitative aspects)	Pertinent (qualitative aspects)

So, in conclusion, the new health information system (HIS), consequent to the Health Reform, with the innovation introduced and the relative functions assigned to the diverse levels of the NHS delineated an "information cycle":

1. With new informative exigencies:
   (a) Of AETIOLOGICAL type (search for the determinants of diseases);
   (b) Of EPIDEMIOLOGICAL type (state of health of the population);
   (c) Of EVALUATIVE type (control of efficacy and efficiency).
2. With different utilisation of the data:
   (a) At the local, district, regional and national levels;
   (b) For cognitive purposes (information for information's sake); but also
   (c) Interpretative (information for intervention);
   (d) For government functions (activity planning) and operative motives (resource management).

The concepts requiring emphasis were:

(a) HIS (Health Information System) as an organic structure for the collection, elaboration and offering of information about mortality, morbidity and, more in general, about the state of health of a population and about the relative health services (structures and their utilisation), with system choices such as: Finality; Communication between levels; Uniqueness, circularity, selectivity and finalisation of the information;
(b) Relations between the Information System and the Organizational System: the Information System should not have constituted a channel differentiated from, separate from, lateral to or superimposed over the Organizational System but rather it should have coincided with the organisation of works and services, with operator activities, with the intervention projects and the social health policy decisions;
(c) Information, then, and not mere data, with the capacity to clearly express the context, the form, the entity and the quality of the knowledge necessary (the information requirement) to be merged with the available at commencement (the information base).

During the 1980s, with actuation of National Statistical Institute (ISTAT) reforms, many of the indications above were applied and a process of integration of the flows and diverse information sources commenced, arriving, in 1989, with the institution of SISTAN, which, in its' triennial National Statistical Programs (NSP) also sought a synthesis between current statistics and ad hoc investigations and opened the way to a system of indicators.

Mortality Statistics and HAA (Hospital Activity Analysis)

The two sectors which, in Italy, have historically distinguished statistical surveys in the health field were those of mortality on one hand and those of hospital admissions on the other.

From the International List of Causes of Death (ISI, Vienna 1891), with the principles enunciated in the 1899 Resolution, through achieving unification of the list of causes of death with that of diseases under the acronym ICD (International Classification of Diseases, Injuries and Causes of Death) in its' periodic (10 yearly) reviews under the aegis of the WHO, causes of death were (and to some extent still are) the base element (due to the characteristic reliability, completeness, etc.) for international comparisons but also for indirect analysis of potential environmental or other risk factors (see territorial mortality maps and cluster analyses).

From the simple mortality rates (specific and standardised) in the 1990s we finally achieved the DALYs (Disability Adjusted Life Years) highlighting the components due to the YLLs (Years of life lost) and the YLDs (Years lost to disability).

Hospital admissions were another point of interest (WHO Survey of Hospital Discharge Summary, 1976) with the proposal of an international minimum common denominator (Minimum Basic Data Set) for hospital discharge forms.

In Italy, after an initial phase of sample survey of those discharged in the first 7 days of the month, now a total survey of hospital admissions provides data for epidemiological purposes and, above all, for administrative/management purposes (see reimbursement procedure based of the DRGs—Diseases Related Groups).

In this context the best known indices were used (average stay in hospital, occupation, rotation, turn-over interval) and the diagrams for synthetic display of internal management performance (Barber) and the relative user population (Gandy).

Brief History of Health Indicators

At the beginning we find the "Movement of Social Indicators" (Moser 1960s) and the 1970 OECD Programme, intended to identify "social demand", in an attempt to "specify and define the concept of social wellbeing with a strictly scientific method which might be assumed and presumed to be universally applicable".

Which "historical" definitions of social indicators might we remember:

Cohen (1970): "It may be defined as a statistic of direct normative interest which facilitates concise, comprehensible judgments balanced on the condition of the major aspects of society. It is, in any case, a direct measurement of wellbeing".

Land (1971): "They are social statistics which: (a) are components of a social system model; (b) can be collected and analysed at different times and accumulated in temporal series; (c) can be aggregated or dis-aggregated at levels adequate to the specification of the model".

CENSIS (1980): "Classical measurements of the development of a nation".

Curatolo (1981): "A quantitative information item about any aspect of reality (with reference to a territory of to a social group) which is useful in clarifying the situation in which that territory or group is with respect to the phenomenon considered from time to time".

While as "historic" definitions of health indicators we refer to:

Balinsky and Berger (1975) "A health indicator is a quantitative series of variables which describe the health conditions of a population".

Ellinson (1977): "those primarily social dimensions of health, applicable not only to individuals but also to large aggregations of individuals (workers, the insured, communities, regions, nations) with a useful measurable nature".

WHO (1980): "The indicators are, substantially, accurately selected information, which assist the measurement of changes relative to priority criteria and enable monitoring of specific aspects of health policy or factors pertinent to the determination of health or health related policy".

Taking note of the evolution of the description towards operativity, we can define the indicators as "information selected for the scope of knowing the phenomena of interest, measuring the changes therein and, consequently, contributing to and orienting the decisional processes at diverse institutional levels" AAVV (1983).

As prototypes of "Lists of indicators" proposed in the past at the international level, we recall, for the two spheres:

(a) List of OECD indicators (1981)
(b) List of useful indicators in the WHO 1980 Programme "Health for All by 2000".

Health Indicators in Italian Regulations

An initial regulatory reference to health indicators is found in Articles 10 and 14 of Legislative Decree No. 502/1992. In particular, Article 10 introduces the method of verification and review of the quality and quantity of services furnished by the NHS and commits the Ministry of Health to define the contents and modality of use of the indicators of efficiency and quality. It establishes in particular:

– For the central level the duty of defining objectives of a national nature and the essential levels of assistance and verification of overall results produced by the NHS, in terms of both efficacy and of the quality of the assistance provided, and the degree of coverage of the levels of assistance;
– For the regional level, the duty of defining the organisational models for services, supporting and guiding companies in management control and the assessment of service quality, verify the levels of assistance effectively guaranteed within their territorial competency;
– For the District level, the duty of assuring the congruency of the levels of assistance furnished with the needs of the population, assessing their productive efficiency and controlling the costs and quality of services procured.

Utilisation of the instrument represented by the indicators for purposes of monitoring health assistance is emphasised by Article 28, comma 10 of Law No. 448/98 (the Stability Pact) with the commitment to define indicators and parameters concerning the structural and organisational aspects of regional health systems and the expenditure levels in order to verify the assistance levels assured in each region, to evaluate the economic-management results and identify the causes of eventual deficits.

In Legislative Decree No. 56/2000, which emanated dispositions in matters of fiscal federalism, Article 9 ordered the timely activation of monitoring procedures for the health assistance effectively furnished in each region and compliance with the guarantees foreseen at Article 1 of Decree No. 502/1992 (personal dignity, health needs, equity, quality, appropriateness, economy). That same Decree, with the definition of a system of guarantees to achieve, in each region, health tutelage objectives pursued by the NHS, foresaw:

(a) A minimum set of indicators and parameters of reference relevant for purposes of monitoring compliance, in each region, with the essential assistance levels;
(b) The rules for surveying, validation and elaboration of the information and statistical data;
(c) The procedures for periodic publication of the results and identification of the regions which do not respect or do not converge towards the parameters of reference.

Finally there is the Ministerial Decree of 12/12/2001 on the "System of guarantees for health assistance monitoring".

The declared purpose was to establish "a minimum set of indicators and parameters of reference addressing monitoring of compliance, in each region, with essential, uniform levels of assistance and with the budgetary constraints of the Regions with ordinary statutes".

The definitions of the indicators, the base data, the parameters of reference, the indicator selection criteria for inclusion in the minimum set, the criteria for aggregation of the indicators on the basis of uniform, essential levels of assistance, for the quality and validation of the base data are indicated below.

The list of indicators, referring to:

(a) Collective health assistance in the living and working environment;
(b) District assistance;
(c) Hospital assistance,

are followed by the result indicators, the state of health indicators, the process quality indicators as well as demographic and socio-economic indices and personal and behavioural factors which affect health and environmental factors linked with living and working conditions.

As many as 91 forms detail, finally, each indicator, providing a definition, the source of the data, the denominator and the value of reference.

The Current Scenario of Indicators Utilised in Health Government

The current situation of usage of the indicators from the Italian Health System is essentially dominated by two factors: the economic situation and federalism.

The economic situation has effectively consumed the attention and efforts of the deciders and planners often leading to underestimating or even ignoring information about the population's needs and health.

Health Indicators

With the exception of sectors dealing with prevention, epidemiological information was left, ultimately, to "the connoisseurs" of the subject and the funds assigned ever less as were research projects to describe the state of public health. Mortality studies, incidence studies, etiological studies today gain increasingly less attention from those governing the Italian health system and are only approached by peripheral services which sometimes perform them in a noteworthy accurate, stimulating manner.

So today there is no growth at the institutional level of availability of purely "epidemiological" indicators capable of providing a snapshot of the presence of pathologies in the Italian population. Despite this the ISTAT mortality data, ISTAT surveys of the state of health, the cancer registers, the data from certain sector groups like the one dealing, for example, with diabetes and others are able to provide interesting indications which, unhappily, are almost entirely ignored in the planning processes.

An interesting estimate of the prevalence of certain pathologies is made available today by the record linkage system like that implemented by the Lombardy Region denominated BDA (Beneficiary Database). This system estimates prevalence utilising algorithms which merge consumption information for different assistance furnished to maximise the sensitivity and specificity of the nosological classification assigned.

Cost and Expenditure Indicators

These are the indicators which gain the most attention from current Italian health managers. The indicators are essentially constructed on two survey systems: the EA system (economic accounts) and the AL system (assistance levels costs).

If the EA indicators traditionally satisfy the needs of accounting control, the AL indicators are more innovative for Health management in that they ought to be able to measure the costs of assistance, at both company and regional level, articulated by level and sub-level.

Obviously for a planning process it is essential to know not just from the EA indicators how much personnel costs, the cost of acquiring goods and services, etc., but also to use the AL indicators to indicate hospitalisation, specialist assistance, emergency, prevention costs, etc.

Utilisation in this sense is foreseen in the Pact for Health (a protocol in the State-Regions Accord of 3 December 2009) which, at Enclosure 2, reports a series of indicators for evaluation. These indicators refer to pro capita and mean costs for services in the divers assistance sectors.

The pro capita costs should enable highlighting the balance of allocation of funds, eventually identifying situations of excess or shortage. The mean costs per single service/intervention should rather have the capacity to estimate economic efficiency as a measurement of the cost of a single service/intervention.

Setting aside the data reliability which has still not achieved the levels desired, the principal problem is that of adjustment of the indicators, for example for risk and/or gravity factors (case-mix). The difficulties of working in this direction also derive from the fact that the Decree on standard costs on health approved recently by the government expects to use these indicators for purposes of determination of the financial requirements of the regions, and so there is stimulation to select calculation methodologies which orient the results in a direction more favourable to one's own "interests".

Fulfilment Indicators

A further sector which has developed is that of the indicators of fulfilment of the obligations on the Regions regarding governance of the regional health systems.

Much of this fulfilment is prevalently bureaucratic in nature, consisting of the production of a specific document: yet the principal interest here is for indicators of satisfaction of levels of assistance.

Twenty-one subjects have been identified (several articulated across multiple indicators) for which regional fulfilment is measured and this is compared with the values of reference which highlight both the differential and by a value considered correct and the region's decreasing or growing distance from that value in the last three-year period considered.

Each of these subjects thus obtains a fulfilment score and each is then assigned a relevance weighting to enable the calculation of a weight sum of the overall value of the indicators. This value indicates whether the region is compliant/virtuous or must adapt to avoid sanctions or is non-compliant and is sanctioned.

This one is perhaps the sole important example of indicators which are inter-synthesised and which are not reduced to a role of mere description but for which there is a rigidly designed decisional mechanism.

A similar system yet to be concluded is that of the so-called indicators of constitutional guarantee of levels of assistance. An initial hypothesis foresaw a list of indicators and articulation thereof which rose to almost 1,000 values. It was clear that such a wide analytical system could not be utilised to perform global evaluation of an entire regional health service, so there is an attempt to articulate it into three levels of indicators: the first with a limited number of indicators whose values may not fall below the guarantee limits, a second contains indicators which highlight shortfalls in the single regions but for which those regions are simply obliged to motivate and explain the situation and describe how they intend resolving the same and finally a third with context indicators useful for the purpose of interpreting and adjusting the first two categories.

Determination of the values of reference is a very serious problem, as sometimes it is preferable to indicate values reported in literature and at other times values referring to the national distribution: in any case the concept of standard values is anything but a simple concept agreed by all.

Indicators of Appropriateness

A further set of indicators which is always evolving is that of the indicators of appropriateness. The concept of appropriateness is articulated in several dimensions, first of all that of organisational appropriateness and clinical appropriateness. The indicators available today refer almost exclusively to organisational appropriateness as it is difficult, if not almost impossible, to evaluate clinical appropriateness with administrative data.

The Pact for Health lists eight indicators of appropriateness which are almost all already calculated and described on the Ministry of Health site. However these indicators contain diverse types of difficulties and are often misleading also with respect to indicators available at the international level with the same name which are however calculated differently.

For example, the percentage indicator of operations for fractures of the femur within 48 h considers: (a) only fractures to the head of the femur and/or to the hip and not extra-articular fractures and on the other hand sometimes they are summed

together; (b) the calculation of the 48 h is technically impossible as they only have the date of admission and of the operation and thus they include operations on the second day and others include the third day of admission; (c) these cannot take into consideration operations in admissions commencing with transfer from another hospital as the calculation should be from the time of the fracture to the time of operation and not the time of admission to the hospital; (d) any case where the general conditions impede the operation should be excluded from the calculation as in the case of grave multi-trauma patients, etc. Depending on how the calculation modalities are specified, the values can vary significantly.

Performance Indicators

Recent administrative legislation increasingly mandates the activates of service and delivery establishment assessment processes. Evaluation of a health cannot but be a multi-factorial evaluation and some consider that it can however be synthesis into a single performance indicator.

Various systems have been produced which are mostly inspired by the Balanced Score Card system; of these, recently the so-called Target system has raised some interest, as it graphically synthesises some thirty indicators which, in turn, synthesise others. These systems raise numerous doubts and criticism:

(a) First of all there is the problem of defining performance. At the intuitive level we can also intuit this but when the responses of different operators and citizens are analysed one notes that each attributes a different meaning to the concept of performance.
(b) It is thus important, first of all, to see which are the dimensions of performance, of which we list four examples, the 4 Es: efficacy, equity, efficiency, economy, while others add other elements.
(c) Having defined the dimensions, it is then important both to see how their measure, may be obtained by synthesising the relative analytical indicators, and to ask oneself whether the dimensions are orthogonal to each other or logically correlated and thus may be synthesised with different methods (in any case for both analytical and indicators of synthesis of several dimensions it is necessary to define the metrics before they are synthesised and if, e.g. one thinks of transforming them all to their standardised "z" values using means and standard deviations of the national distributions, then it is necessary to establish whether the synthetic value weights or not each single indicator to the same measure).

This type of system, abroad, has led to the publication of league tables much criticised in the methodological profile (punctual values should be supported by the relative confidence intervals!) and because they often tend to orient service users in a distorted manner.

Outcome Indicators

The so-called outcome indicators are the last series and they seek to estimate the efficacy of treatments, especially but not exclusively surgical treatments.

The indicator used most is that of mortality at 30 days of the disease event or operation. The greatest problem with these indicators is surely that of the risk adjustment cannot yet be calculated on clinical analytical data although it can on the little information available in the data collected for mainly administrative purposes, such as that of the HDOs (the current hospital discharge sheets).

Themes and Criticalities in the Definition, Calculation and Use of Health Indicators

Here below, in concluding this brief excursus, we provide a list of themes which constitute the arguments for current discussion in matters of health indicators; there is insufficient space for a thorough review but sufficient for a mention of the diverse problems.

Use of the Indicator in Planning and Decisional Processes

From Opinion and Intuition to Measurement and Evidence

There is increasing recognition of the need to base decisions on measurements and not just on opinions albeit from experts. Planning cannot rely on pseudo-certainties not based on empirical evidence.

Indicators as a Basis and Indicators as a Justification of Prior Decisions

Indicators are used more frequently to justify decision taken on the basis of other criteria than to provide the basis therefore.

The Ex-Ante, During and Ex-Post Indicator in Decisional Processes

Indicators have a role which is both preparatory to decisions, and of verification during the implementation process and, finally, of evaluation of output and outcomes.

Subjectivity, Impartiality and Subjectivism in the Production of Indicators

The distinction between subjectivity and subjectivism is highly relevant: identification of the elements of subjectivity of users is very important but the elements of subjectivism in survey and interpretation processes must be controlled.

From the Fact to the Datum: Quality and Organisation of Information Systems

Interaction Between Decisional Interests and Correctness of Information

The information system must of itself not be influenced by decisional interests which may alter the veracity thereof; it is true that the quality of information improves proportional to the interest one has in having that information.

Data for Administrative Management and Planning Governance Purposes

In the main the data necessary to administrative management is more complete than collected for information purposes only.

Private Confidentiality and Public Accessibility of the Data

If respect for privacy must be considered an important element, the same is also true for respect for the public interest seeking, in any case, to guarantee the maximum possible level of confidentiality.

Quality Problems: Sampling Errors, Survey Bias, Distortion

There is often a fear of sampling "errors" while by contrast far more serious distortions due to the survey systems are underestimated.

From the Datum to the Information: Indices, Indicators, Valorisations, Standard, Target

Correctness and Transparency of the Algorithms

The decider must be able to at least intuit how the indicator calculation algorithm works, an algorithm which must obviously have a scientifically based correctness.

From Description to Evaluation: The Functions of Valorisation

This point is essential: the indicator describes a variability of the reality but measurement of the indicators must be associated with value assigned thereto.

The Question of Risk Adjustment

Correction for the confusing covariance enables the highlighting of the elements which constitute the basis of the indicator assignable to determinant component thereof.

Which Value of Reference: Standard and/or Target

The standard is an element which should be objective and as such is often difficult to determine; the target on the other hand corresponds only to an objective which the decider has set for a determined activity.

From Mono-Dimensional Indicators to Complex Scenarios

The Function of Composition of Elementary Indicators

The passage from an elementary indicator to a complex indicators requires the application of a composition function which, on the one hand does not distort the reality yet on the other manages to hold the principal elements of the complexity itself.

The Degree to Which the Information Can Be Synthesised

It is often better to abstain from synthesis operations on elementary indicators leaving this task to the user of the indicators.

The Debatableness Quotient in the Fusion of Inter-Orthogonal Dimensions

The logical sum of different information dimensions requires the application of weightings which are absolutely subjective and debatable; it is therefore opportune that the decider defines those weightings.

Distortions in Processes of Communication of Complex Scenarios

The phase of communication of the syntheses of an information process can also imply notable distortions which thus must be avoided evaluating the mechanisms on which these may be generated.

From the Scenarios to Policies

Indicative or Automatic Utilisation of the Information in Decisional Processes

It is opportune that there be indicators which automatically activate decisions and indicators which rather have a merely indicative role for those taking the decisions.

Indicators for Vertical Decisions or for Horizontal Evaluations and Comparisons

Vertical decisions are between hierarchical levels in the same system while horizontal decisions are those which refer to the homogeneity between levels in different systems.

Cost–Benefit in the Production of Information Scenarios

There must always be a favourable ratio between the cost of a system of indicators and the usefulness of that system; costly systems of little use must, absolutely, be banished.

Communicability Between Technicians and Politicians in Their Respective Fields of Expertise

The system must be comprehensible to politicians and the technicians must be able to identify the elements which must be expressed by the politicians.

Improving the Value of the Reality or Improving the Value of the Indicator

One risk to be avoided is that of activating policies to improve the indicators without considering improvement of the reality: the indicators are a representation of the reality and are not the reality itself!

References

AAVV: Gli indicatori sanitari, Atti del Seminario su "Gli indicatori Sanitari", Bologna, 27–28 Novembre 1981, in Epidemiologia e Prevenzione, Monografia n° 19–20, Anno VI (1983)

Marchi, M., Cislaghi, C., Bottasso, F.: Statistiche Sanitarie Correnti e Osservazione Epidemiologica. Epidemiologia e Prevenzione **4**, 18–24 (1978)

OECD: O.E.C.D.'s social indicators: a measure of well-being. OECD Observer **113**, 32 (1981)

WHO: The use of health indicators – How to make statistics talk, Report on a WHO Workshop, Brussels, 10–14 November (1980)

Measuring the Multidimensional Demographic Convergence by Indices of Multiple Variability

Maria Rita Sebastiani

Abstract

We aim to test the demographic convergence of the populations towards a common pattern. Many studies have used different statistical indices of variability, but they have only focused on a one-dimensional perspective, testing the convergence of each variable at time. In fact, each population is characterized by different demographic phenomena; then it could be interesting to test the convergence in a multidimensional perspective, considering all variables together at one time. We propose a statistical method useful for this aim. We consider the crude birth rate, the crude death rate, the infant mortality rate, and the aging index. We define suitable absolute indices of multiple variability and, in an aim to evaluate the degree of the convergence, the corresponding normalized ones with values comprised between 0 and 1. We test the convergence of the European populations for all years from 1960 to 2007.

Introduction

We aim to measure the demographic convergence of the populations towards a common pattern. Many studies have already used different approaches and methods. For instance, in an aim to test the convergence of fertility, Crenshaw (Crenshaw et al. 2000) estimated a regression model on the growth rate of the fertility rate, where the coefficient β measures the degree of the convergence (β-*convergence*). Other researchers have used statistical indices of variability such as, for instance, the standard deviation, the variation coefficient, or the range of variation (σ-*convergence*):

M.R. Sebastiani (✉)
Department of Methods and Models for Economics, Territory and Finance, Sapienza Università di Roma, Roma, Italy
e-mail: mariarita.sebastiani@uniroma1.it

F. Crescenzi and S. Mignani (eds.), *Statistical Methods and Applications from a Historical Perspective*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-05552-7_35,

if the variability is low, it means that the populations converge towards a common pattern; if the variability is high, the populations are quite different from each other and the convergence is not achieved. Recently, we proposed an absolute index of variability and the corresponding normalized one (by symbols, respectively, S_{jj} and $S_{jj,\text{norm}}$) that seem to us the most suitable indices when the variable is a demographic rate (Sebastiani 2010). However, all the existing studies have focused on a one-dimensional perspective, testing the convergence of each variable at time. In fact, each population is characterized by different demographic phenomena; then it could be interesting to test the convergence in a multidimensional perspective, considering all variables together at one time. Indeed, populations that are similar to each other with respect to some variables (for instance, in terms of mortality) could differ with respect to others (for instance, in terms of births), meaning that they are in different stages of transition. Here, we propose some statistical indices of multiple variability useful for this aim.

Let us consider a set of k variables, each observed on n populations. Since the demographic transition theory concerns with changes of births, mortality, and age structure (Thompson 1929; Landry 1934; Notestein 1945), we consider the crude birth rate, the crude death rate, the infant mortality rate, and the aging index (by symbols, CBR, CDR, IMR, and AI). Since CBR, CDR, IMR, and AI are statistical rates, we define suitably the variance and covariance matrix $\mathbf{S}$. Specifically, the variance has the expression of S_{jj}, whereas the covariance is defined similarly to S_{jj}. Since the rates can differ from each other with respect to the magnitude (CBR, CDR, and IMR take values comprised between 0 and 1, whereas AI can take also values greater than 1) and, consequentially, with respect to the degree of variability, it seems suitable to consider also the correlation matrix $\mathbf{R}$ defined from $\mathbf{S}$.

For summarizing the variability of the n k-dimensional populations, we propose three absolute indices of multiple variability: the trace of $\mathbf{S}$, the determinant of $\mathbf{S}$, and the determinant of $\mathbf{R}$ (by symbols, respectively, $\text{tr}(\mathbf{S})$, $\det(\mathbf{S})$, and $\det(\mathbf{R})$). Since $\text{tr}(\mathbf{S})$ is the sum of the variances of the k rates, we expect that $\text{tr}(\mathbf{S})$ reproduces the values of S_{jj} applied to each of the k rates at time. Instead, we expect that $\det(\mathbf{S})$ and $\det(\mathbf{R})$ can highlight eventual lags among stages of transition of the populations. In an aim to evaluate how high is the degree of the convergence, we need to use normalized indices with values comprised between 0 and 1 (where 0 and 1 mean, respectively, maximum convergence and maximum divergence). For this reason, we apply a linear normalization procedure to each absolute index, after determining for each of them both the minimum and maximum values, and so we obtain the corresponding normalized ones (by symbols, $\text{tr}(\mathbf{S})_{\text{norm}}$, $\det(\mathbf{S})_{\text{norm}}$, and $\det(\mathbf{R})_{\text{norm}}$).

We use the normalized indices for testing the demographic convergence of the populations of the European Union (EU). We consider six different frameworks of the EU according to the successive expansions (specifically, with 6,[1] 10,[2] 12,[3]

[1] Belgium, France, German Federal Republic, Italy, Luxemburg, Netherland.

[2] EU6 and Denmark, Greece, Ireland, United Kingdom.

[3] EU10 and Portugal, Spain.

15,[4] 25[5], and 27[6] members) and the six corresponding groups of populations (by symbols, EU6, EU10, EU12, EU15, EU25, and EU27). We measure the multiple variability within each group across time: decreasing trend means progressive convergence, whereas stationarity or increasing trend indicates delays or reversals in the convergence. Comparing the values and trend that a normalized index takes within the different groups, one can identify the most and the less converging groups and, moreover, the time where eventually occur lags among the processes of convergence of the different groups.

The chapter is so organized. In Sect. 2, we describe the data. In Sect. 3, we introduce the matrices **S** and **R**, and define the indices tr(**S**), det(**S**), and det(**R**) and $\mathrm{tr}(\mathbf{S})_{\mathrm{norm}}$, $\det(\mathbf{S})_{\mathrm{norm}}$, and $\det(\mathbf{R})_{\mathrm{norm}}$. In Sect. 4, we discuss the results obtained testing the convergence of the EU's populations. Section 5 contains some concluding remarks.

Data

For each of the 27 EU's populations we observe the yearly series of CBR, CDR, IMR, and AI, for all the years from 1960 to 2007. Each yearly series describes an individual demographic profile. For methodological issues concerning with the normalization of the absolute indices, we need to take the data about the stocks and flows that are used for calculating the rates. In choosing the source of data, we met some difficulties due to the unavailability of detailed electronic data simultaneously for all the rates into one source. Therefore, for CBR, CDR, and IMR, for the years up to 2004 we take the data from the Council of Europe (Council of Europe eds. 2006), whereas for the years since 2005 from the United Nations (United Nations eds. 2009). For AI, the only source is Eurostat (Eurostat). For each rate, all the individual profiles are quite similar to each other, particularly for CBR and IMR. In particular, there is a very strong similarity among the individual profiles of the first four groups (Central and Western Europe). Instead, sometimes the profiles of the Eastern populations differ from each other as well as compared with those of the other European populations (particularly for CDR, IMR, and AI). These differences could be due to some changes of the lifestyle that occurred after some exceptional events (for instance, the pulling down of the Berlin Wall and the end of the Soviet Union). However, it is very likely that the individual profiles of the Eastern populations are going to be similar to those of the other European populations.

For each rate, we determine the six European common profiles, calculating the average rate for each group: for every year we divide the number of the demographic occurrences (namely, births, deaths, or number of elderly people) by the overall

[4] EU12 and Austria, Finland, Sweden.

[5] EU15 and Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovak Republic, Slovenia.

[6] EU25 and Bulgaria, Romania.

reference population. The six common profiles tend to be adjacent or even to overlap each other. The common profiles of CBR, CDR, and IMR decrease, whereas AI increases notably.

In (Sebastiani 2010) we discussed the results obtained applying $S_{jj,\text{norm}}$ to CBR, CDR, IMR, and AI in an aim to test the one-dimensional convergence of the EU's populations. They converge towards a common pattern, which is typical of a "mature" population with low natural increase and high aging. Indeed, for each rate and within each group, variability is always low; moreover, for CDR and IMR, it decreases across time. Instead, for CBR and AI it slightly rises in the final period (more markedly in EU6 and EU10), meaning that there are some weak reversals in convergence. For CBR this is due to the increasing births in some very large populations (for instance, in the French population for effect of some demographic policies as well as of the higher fertility of the immigrant women).

Indices of Multiple Variability

Let us use a general notation. Let $S = \{1, \ldots, i, \ldots, n\}$ and $V = \{1, \ldots, j, \ldots, k\}$ be two sets of labels representing, respectively, n populations belonging to a same group and k rates (where n and k are fixed integer numbers). Specifically, here $n = 6$, 10, 12, 15, 25, or 27, according to the group. Let $(t_{h1}, \ldots, t_{hj}, \ldots, t_{hk})$ be the ordered set of the k rates values observed on the hth population of S ($1 \le h \le n$), with $t_{hj} = x_{hj}/p_{hj}$, where x_{hj} and p_{hj} are, respectively, the variable (for instance, the number of births, deaths, or elderly people) and the reference population for x_{hj}. Let $\bar{t}_j$ be the (weighted) average rate of the n rates t_{hj} over S, that is: $\bar{t}_j = \dfrac{\sum_{h=1}^{n} t_{hj} p_{hj}}{\sum_{h=1}^{n} p_{hj}} = \dfrac{\sum_{h=1}^{n} x_{hj}}{\sum_{h=1}^{n} p_{hj}}$ $(1 \le j \le k)$.

Let $\mathbf{S} = \{S_i\} \in \mathrm{M}_{kk}$ be the variance and covariance matrix of the k rates. Moreover, let $\mathbf{R} = \{R_{ij}\} \in \mathrm{M}_{kk}$ be the correlation matrix of the k rates obtained from $\mathbf{S}$. Since $\bar{t}_j$ are weighted averages, we believe that it is right to take also the variance and the covariance of the rates as weighted averages. We define:

$$S_{jj} = \sum_{h=1}^{n} \left[\left(t_{hj} - \bar{t}_j\right)^2 \frac{p_{hj}}{\sum_{h=1}^{n} p_{hj}} \right] \tag{1}$$

$$S_{ij} = \sum_{h=1}^{n} \left[\left(t_{hi} - \bar{t}_i\right) \left(t_{hj} - \bar{t}_j\right) \sqrt{\frac{p_{hi}}{\sum_{h=1}^{n} p_{hi}} \frac{p_{hj}}{\sum_{h=1}^{n} p_{hj}}} \right] \tag{2}$$

where S_{jj} is the same index already introduced in Sebastiani (2010).

Obviously, R_{ij} is defined as: $R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$.

In an aim to measure the multiple variability of the n populations, we summarize the matrices $\mathbf{S}$ and $\mathbf{R}$ taking tr($\mathbf{S}$), det($\mathbf{S}$), and det($\mathbf{R}$). Specifically, tr($\mathbf{S}$) measures the (weighted) average distance of the n k-dimensional populations from the common profile $\left(\bar{t}_1, \ldots, \bar{t}_j, \ldots, \bar{t}_k\right)$, assuming that the k rates are independent. Instead, both det($\mathbf{S}$) and det($\mathbf{R}$) measure the volume of the k-dimensional space where are represented the n populations. These two indices seem more adequate than tr($\mathbf{S}$) for measuring the distance between the n populations when the k rates are not independent, since they take into account of the correlation between the rates.

For normalizing tr($\mathbf{S}$), det($\mathbf{S}$), and det($\mathbf{R}$), we determine for each of them both the minimum and maximum values and apply a linear normalization procedure. In particular, since both $\mathbf{S}$ and $\mathbf{R}$ are positively or semi-positively defined, we have that:

$$0 \leq \operatorname{tr}(\mathbf{S}) \leq \sum_{j=1}^{k} max\left(S_{jj}\right) \Rightarrow \operatorname{tr}(\mathbf{S})_{\text{norm}} = \frac{\operatorname{tr}(\mathbf{S})}{\sum_{j=1}^{k} \max\left(S_{jj}\right)} \tag{3}$$

$$0 \leq \det(\mathbf{S}) \leq \left(\frac{1}{k}\sum_{j=1}^{k} \max\left(S_{jj}\right)\right)^{k} \Rightarrow \det(\mathbf{S})_{\text{norm}} = \frac{\det(\mathbf{S})}{\left(\frac{1}{k}\sum_{j=1}^{k} \max\left(S_{jj}\right)\right)^{k}} \tag{4}$$

$$0 \leq \det(\mathbf{R}) \leq 1 \Rightarrow \det(\mathbf{R})_{\text{norm}} \equiv \det(\mathbf{R}) \tag{5}$$

where max(S_{jj}) represents the maximum value of S_{jj} ($1 \leq j \leq k$). For determining numerically the values of tr($\mathbf{S}$)$_{\text{norm}}$, det($\mathbf{S}$)$_{\text{norm}}$, and det($\mathbf{R}$)$_{\text{norm}}$, we implement an algorithm in R language (R Development Core Team 2009).

Results

In applying the proposed methodology, we take several subsets of k rates selected among CBR, CDR, IMR, and AI, with $k \geq 2$ (since we deal with multiple variability). Indeed, we expect that applying a normalized index to different sets of rates may obtain different values and trends. For instance, variability could decrease with an increase in k due to the correlation between the rates. Moreover, variability calculated for a set of rates that include IMR is expected to be on average higher than for other sets that do not include IMR, due to the fact that on average IMR shows higher variability than the other rates. Variability calculated for a set of rates that include AI could be on average lower than for other sets that do not include AI, due to the fact that on average AI shows lower variability than the other rates. Since

Table 1 $\text{tr}(\mathbf{S})_{\text{norm}}$ by set of rates and group of populations: mean values

	$k=2$ (CBR, CDR)	$k=3$ (CBR, CDR, IMR)	$k=3$ (CBR, CDR, AI)	$k=4$ (CBR, CDR, IMR, AI)
EU6	0.00067	0.00105	0.00030	0.00030
EU10	0.00021	0.00047	0.00020	0.00020
EU12	0.00026	0.00089	0.00019	0.00019
EU15	0.00025	0.00084	0.00018	0.00018
EU25	0.00029	0.00108	0.00016	0.00016
EU27	0.00030	0.00142	0.00016	0.00016

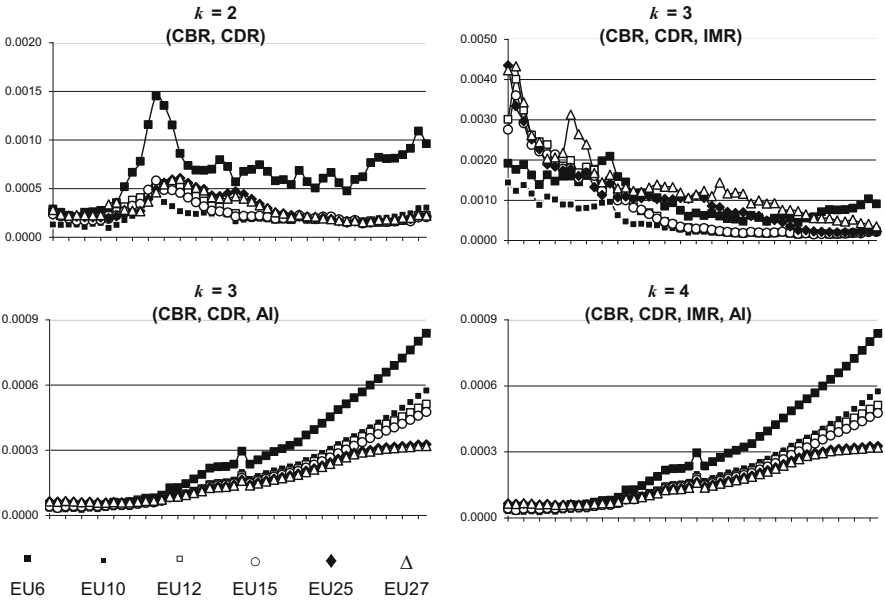

Fig. 1 $\text{tr}(\mathbf{S})_{\text{norm}}$ values by set of rates and group of populations

the demographic transition theory refers to the simultaneous changes of births and mortality, we take the only sets that include both CBR and CDR. Thus, for $k=2$ we consider (CBR, CDR), whereas for $k=3$ take (CBR, CDR, IMR) and (CBR, CDR, AI). For each set of rates, and for each group of populations, we construct the matrices $\mathbf{S}$ and $\mathbf{R}$, calculate $\text{tr}(\mathbf{S})_{\text{norm}}$, $\det(\mathbf{S})_{\text{norm}}$, and $\det(\mathbf{R})_{\text{norm}}$, and study the pattern of these indices across time.

Table 1 and Fig. 1 show the mean values and the pattern of $\text{tr}(\mathbf{S})_{\text{norm}}$ across time.

The results show that the populations converge towards a common profile: indeed, $\text{tr}(\mathbf{S})_{\text{norm}}$ takes so small values that, for highlighting the differences of variability of the groups, we must take the values to five decimal places. We attain some results that we have expected. For each set of k rates, the pattern of $\text{tr}(\mathbf{S})_{\text{norm}}$ is the resultant of the pattern of S_{jj} applied to each of the k rates at time. For each group

Table 2 $\det(\mathbf{S})_{\text{norm}}$ by set of rates and group of populations (tenth root transformations): mean values

	$k=2$ (CBR, CDR)	$k=3$ (CBR, CDR, IMR)	$k=3$ (CBR, CDR, AI)	$k=4$ (CBR, CDR, IMR, AI)
EU6	0.196	0.085	0.007	0.001
EU10	0.169	0.075	0.008	0.002
EU12	0.177	0.082	0.008	0.002
EU15	0.176	0.082	0.009	0.002
EU25	0.186	0.090	0.009	0.002
EU27	0.187	0.096	0.009	0.002

of populations, $\text{tr}(\mathbf{S})_{\text{norm}}$ calculated for (CBR, CDR, IMR) is on average slightly higher than for the other sets of rates; moreover, $\text{tr}(\mathbf{S})_{\text{norm}}$ calculated for (CBR, CDR, AI) is on average slightly lower than for the other sets of rates. On average, for each group, $\text{tr}(\mathbf{S})_{\text{norm}}$ calculated for the set with all four rates is identical to the values for (CBR, CDR, AI). Let us consider the pattern of $\text{tr}(\mathbf{S})_{\text{norm}}$ across time. Variability calculated for (CBR, CDR) shows the same pattern of $S_{jj,\text{norm}}$ calculated for CBR, due to the higher variability of CBR than CDR: $\text{tr}(\mathbf{S})_{\text{norm}}$ decreases in all groups except for EU6 and EU10. EU6 shows on average the highest variability. Considering (CBR, CDR, IMR), the more intense fall of variability of IMR than both CBR and CDR makes $\text{tr}(\mathbf{S})_{\text{norm}}$ decrease within all groups (less markedly in EU6). EU27 shows on average the highest $\text{tr}(\mathbf{S})_{\text{norm}}$ values. Considering (CBR, CDR, AI) and (CBR, CDR, IMR, AI), the notably increasing variability of AI makes $\text{tr}(\mathbf{S})_{\text{norm}}$ increase within all groups, particularly within EU6 (that shows the highest values since 1974, due to the marked aging of the Italian and German populations).

Table 2 and Fig. 2 show the mean values and the pattern of $\det(\mathbf{S})_{\text{norm}}$ across time.

The results show that the demographic convergence is achieved. $\det(\mathbf{S})_{\text{norm}}$ takes such small values that, for appreciating the differences of variability of the groups, we must take a tenth root transformation of the values. Compared with $\text{tr}(\mathbf{S})_{\text{norm}}$, we attain also other important results. For each group, on average $\det(\mathbf{S})_{\text{norm}}$ decreases with k increasing, probably due to the correlation between the rates. For $k=3$, $\det(\mathbf{S})_{\text{norm}}$ calculated for (CBR, CDR, IMR) is on average higher than for (CBR, CDR, AI), as observed also for $\text{tr}(\mathbf{S})_{\text{norm}}$. Moreover, for each set of rates, on average $\det(\mathbf{S})_{\text{norm}}$ increases with n, except for (CBR, CDR) and (CBR, CDR, IMR) where EU6 shows higher values than EU10.

Considering the pattern of $\det(\mathbf{S})_{\text{norm}}$ across time, for each set of rates and for each group, it tends to decrease except for (CBR, CDR), where variability within EU6 and within EU10 increases, as observed also for $\text{tr}(\mathbf{S})_{\text{norm}}$.

Table 3 and Fig. 3 show the mean values and the pattern of $\det(\mathbf{R})_{\text{norm}}$ across time. Values are not always so small: within some groups the individual profiles differ notably from the common profile and therefore the convergence is not achieved. Specifically, calculating $\det(\mathbf{R})_{\text{norm}}$ for (CBR, CDR) and (CBR, CDR, IMR), for all groups we obtain values that on average are relatively high (particularly within

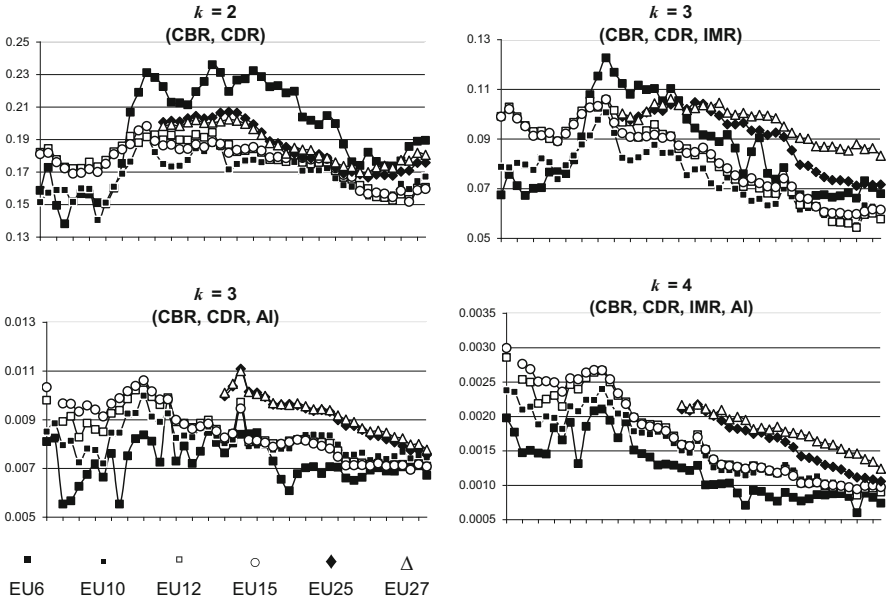

Fig. 2 $\det(\mathbf{S})_{\text{norm}}$ values by set of rates and group of populations (tenth root transformations)

Table 3 $\det(\mathbf{R})_{\text{norm}}$ by set of rates and group of populations: mean values

	$k=2$ (CBR, CDR)	$k=3$ (CBR, CDR, IMR)	$k=3$ (CBR, CDR, AI)	$k=4$ (CBR, CDR, IMR, AI)
EU6	0.394	0.180	0.023	0.006
EU10	0.706	0.597	0.123	0.089
EU12	0.711	0.584	0.122	0.086
EU15	0.739	0.617	0.150	0.109
EU25	0.863	0.631	0.515	0.168
EU27	0.864	0.543	0.531	0.187

EU25 and EU27). As observed also for $\det(\mathbf{S})_{\text{norm}}$, for each group on average $\det(\mathbf{R})_{\text{norm}}$ decreases with k increasing, probably due to the correlation between the rates. Moreover, for each set of rates, on average $\det(\mathbf{R})_{\text{norm}}$ increases markedly with n, showing that there are relevant differences between the individual profiles of the Eastern populations and those of the other European populations.

Considering the pattern of $\det(\mathbf{R})_{\text{norm}}$ across time, for each set of rates, the six curves corresponding to the different groups are sometimes lagged each other, indicating that there are evident lags between the processes of convergence of the groups (particularly for (CBR, CDR) and (CBR, CDR, IMR)). In general, for each set of rates and for each group of populations, $\det(\mathbf{R})_{\text{norm}}$ decreases, except for (CBR, CDR) where within EU25 and within EU27 variability remains stationary to high values, indicating that the demographic convergence is still far away.

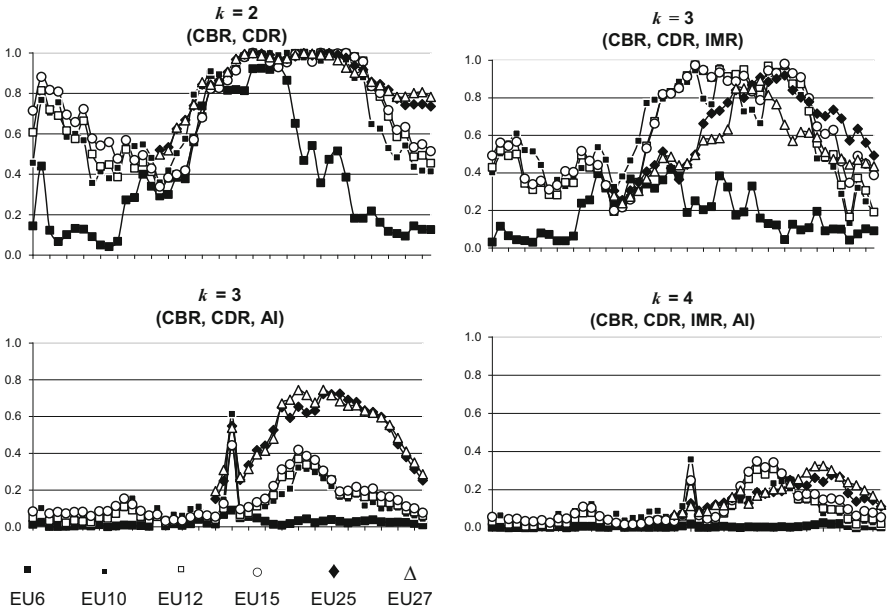

Fig. 3 $\det(\mathbf{R})_{\text{norm}}$ values by set of rates and group of populations

Concluding Remarks

Applying the normalized indices of multiple variability, one can express a precise assessment whether the demographic transition theory is fulfilled and, moreover, how much high is eventually the degree of the convergence. This is the progress with respect to measuring the variability of each of the k rates at time. Anyway, the results obtained applying $S_{jj,\text{norm}}$ to each of the k rates can be useful to understand values of the multiple variability. For instance, the fact that IMR shows relatively high variability can contribute to make raise also the indices of multiple variability; analogously, the fact that AI shows relatively low variability can contribute to make reduce also the indices of multiple variability. In applying the proposed methodology, we take several subsets of k rates selected among CBR, CDR, IMR, and AI, with $k \geq 2$. Indeed, we expect that applying a normalized index to different sets of rates may obtain different values and trends. And so it occurs. Since the demographic transition theory refers to the simultaneous changes of births and mortality, we take the only sets that include both CBR and CDR. Thus, for $k = 2$ we consider (CBR, CDR), whereas for $k = 3$ take (CBR, CDR, IMR) and (CBR, CDR, AI). The results obtained applying the indices to the EU's populations allow to draw some considerations.

As we expected, values of $\text{tr}(\mathbf{S})_{\text{norm}}$ are the resultant of the values of S_{jj} applied to each of the k rates at time. In determining the magnitude and the trend of $\text{tr}(\mathbf{S})_{\text{norm}}$,

the rate with the highest variability seems to have the highest weight. Namely, in calculating $\text{tr}(\mathbf{S})_{\text{norm}}$ for (CBR, CDR) prevails the higher variability of CBR; for (CBR, CDR, IMR) predominates IMR and for the sets that include AI prevails AI.

Compared with $\text{tr}(\mathbf{S})_{\text{norm}}$, the indices $\det(\mathbf{S})_{\text{norm}}$ and $\det(\mathbf{R})_{\text{norm}}$ highlight the higher variability of the groups that are likely the less homogeneous. EU25 and EU27 comprise the Eastern populations, which are in a younger demographic transition stage than the other European populations. These two groups show on average the highest values of $\det(\mathbf{S})_{\text{norm}}$ and $\det(\mathbf{R})_{\text{norm}}$. Instead, applying $\text{tr}(\mathbf{S})_{\text{norm}}$, EU25, and EU27 show on average lower variability than EU6.

Considering the pattern of $\det(\mathbf{R})_{\text{norm}}$ across time, for all sets of rates, the six curves corresponding to the different groups often show different trend, being at a given time some decreasing and others increasing or stationary. Instead, applying $\text{tr}(\mathbf{S})_{\text{norm}}$ and $\det(\mathbf{S})_{\text{norm}}$, for all sets of rates, the curves associated with the different groups show approximately the same trend although with different magnitude. It seems that by means of $\det(\mathbf{R})_{\text{norm}}$, one can highlight also the lags existing among the processes of convergence of the different groups. For instance, for (CBR, CDR), since 1990 variability within EU6 rapidly decreases (meaning progressive and fast convergence), whereas within EU10, EU12, and EU15 remains high and stationary up to 2000 and then decreases (meaning a delay in the convergence); finally, within EU25 and EU27, $\det(\mathbf{R})_{\text{norm}}$ takes constantly the maximum values up to 2000, then decreases slightly up to 2004 and afterwards remains at high values (meaning that the convergence is still far away).

Applying $\det(\mathbf{S})_{\text{norm}}$ and $\det(\mathbf{R})_{\text{norm}}$, for all groups, on average the values decrease with an increase in k, probably due to the correlation between the rates. There are two possible solutions for overcoming this drawback: (a) to measure the convergence by means of a Mahalanobis generalized distance among the profiles, but there could be some problems in inverting the variance and covariance matrix if the rates were perfectly correlated; (b) to apply a suitable statistical technique to the original dataset for defining new uncorrelated variables to use for testing the convergence.

The results obtained applying $\text{tr}(\mathbf{S})_{\text{norm}}$ and $\det(\mathbf{S})_{\text{norm}}$ indicate that the EU's populations converge towards a common profile, which is typical of a "mature" population with low natural increase and high aging. Values of $\det(\mathbf{R})_{\text{norm}}$ show that in some cases the convergence is not achieved and, moreover, that in all groups in the last years there are some reversals and delays in convergence.

Future work will concern with some statistical methodological developments in an aim to include other demographic variables such as the total fertility rate and the life expectancy at the birth (by symbols, TFR and LE). In fact, they could be more appropriate than CBR and CDR for studying reproduction and survival since are not influenced from the age structure of the population, as instead it is the case for CBR and CDR. However, here we could not take TFR and LE in cause of some problems concerning with the normalization of the absolute indices. Moreover, we will study the distributional properties of the indices proposed so that they can be used also for statistical inference.

Acknowledgments I am very grateful to Professor Alessandra De Rose for her useful suggestions that led to a significative improvement in the presentation of the paper. I would like to thank also Professor Raimondo Cagiano de Azevedo for introducing me to the research area of the demographic convergence.

References

Council of Europe eds.: Recent demographics developments in Europe 2005. Strasbourg (2006)
Crenshaw, E.M., Christenson, M., Oakey, D.R.: Demographic transition in ecological focus. Am. Soc. Rev. **65**, 371–391 (2000)
Landry, A.: La Révolution Démographique. Paris (1934)
Notestein, F.: Population: the long view. In: Schultz, T. (ed.) Food for the World. pp. 36–57. Chicago (1945)
R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing. Wien, ISBN 3-900051-07-0. http://www.R-project.org (2009)
Sebastiani, M.R.: Measuring the demographic convergence of the European populations by a normalized variability index. Statistica Applicazioni **2**, 181–198 (2010)
Thompson, W.S.: Population. Am. J. Soc. **34**, 959–975 (1929)
United Nations eds. (2009). Demographic Yearbook 2007. New York (2007)