Cristina Davino
Luigi Fabbris *Editors*

Survey Data Collection and Integration

Springer

Survey Data Collection and Integration

Cristina Davino · Luigi Fabbris
Editors

Survey Data Collection and Integration

Springer

Editors
Cristina Davino
Department of Studies for Economic Development
University of Macerata
Macerata
Italy

Luigi Fabbris
Department of Statistical Science
University of Padua
Padua
Italy

ISBN 978-3-642-21307-6 ISBN 978-3-642-21308-3 (eBook)
DOI 10.1007/978-3-642-21308-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012943361

Printed on acid-free paper

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

Preface

Surveys are an important source of scientific knowledge and a valid decision-support tool in many fields, from social studies to economics, market research, health studies, and others. Scientists have investigated most of the methodological issues on statistical surveys so that the scientific literature offers excellent references concerning all the required steps for planning and realising surveys. Nevertheless, real problems of statistical surveys often require practical solutions that either deviate from the consolidated methodology or do not have a specific solution at all.

Information and communication technology and techniques are earmarking the modern world, changing people's behavior and mentality. Innovations in remote and wireless communication devices are being created at full steam. Purchase behaviors, service fruition, production activities, social and natural events are recorded in real time and stored in huge data warehouses. Warehouses are being created not only for data but also for textual documents, audio files, images, and video files. Survey methods must evolve accordingly: surveys have to take advantage of all new information chances and adapt, if needed, the basic methodology to the ways people use to communicate, memorize, and screen the information. This book is aimed at focusing on new topics in today's survey researchers' agenda.

Let us provide an example of new survey needs. Complex survey designs, which were imagined as statistical means for collecting data from samples of dwellings drawn from population registers, are going to be obsolete but for some official statistics and for rare, relevant, opinion surveys. In fact, the access to population registers has proven strict and face-to-face interviewing has become very expensive. Instead, there are growing possibilities for both private and public organisation for achieving administrative databases related to large population segments. Hence, sample surveys are on the verge of specialisation in collecting subjective information through remote technological devices from samples mechanically drawn from available databases. Our vision is purposively drastic because we want to stress the irreversibility of the trend.

Any survey researcher has to ponder practical and methodological problems when choosing the appropriate technique for acquiring the data to analyse. No researcher is allowed to ignore costs, time, and organisation constraints of data collection. See also the introductory paper by Luigi Biggeri on this issue. This induces a researcher to consider appropriate for her/his purposes:

- Integrating large datasets from various sources rather than gathering ad hoc data by means of traditional surveys;
- Collecting low-cost, real-time massive data rather than gathering data through refined statistical techniques at the expenses of timeliness, budget saving, and respondent bothering;
- Estimating some parameters without sampling error, because all units can be processed, and other parameters based on models that hypothesize at the very local level relationships that are observed at a higher scale, as in the case of small area estimation;
- Connecting and harmonising in a holistic approach, data collected with or analysed at different scales, for instance analyse together data on individual graduates, graduates' families, professors, study programmes, educational institutions, production companies, and local authorities, all aimed at describing the complex (i.e. multidimensional and hierarchical) relationships between graduates' education and work.

Of course, we go nowhere if our research aims are unclear. Supposing aims are clear, then the appropriateness of research choices is a methodological concern. In this volume, a paper by Tomàs Aluja-Banet, Josep Daunis-i-Estadella, and Yan Hong Chen, as well as another by Diego Bellisai, Stefania Fivizzani, and Marina Sorrentino concern the issue of data integration. The first paper deals with the issues to be considered when imputing, in place of a missing value, values drawn from related sources. It deals also with the choice of an imputation model and with the possibility to combine parametric and non-parametric imputation models. The second paper applies to the design of an official survey aimed at integrating data collected on job vacancy and hours worked with other Istat business surveys. A third paper by Monica Pratesi, Caterina Giusti, and Stefano Marchetti discusses the use of small area estimation methods to measure the incidence of poverty at the sub-regional level using EU-SILC sample data.

The very availability of large databases requires researchers to focus on data quality. We mean that data quality assessment is an important issue in any survey, but it tends to the fore just because the sampling error vanishes. This, in turn, requires researchers to learn:

- How to check the likelihood of the data drawn from massive databases whose records could contain errors?
- How to measure the quality of, and possibly adjust, the data collected with opinion surveys that, in general, are carried out by means of high-performance technological tools and by specialized personnel who interview samples of respondents?

- How to guarantee sufficient methodological standards in data collection from key witnesses whom applied researchers turn to more and more often so to corroborate the results of analyses on inaccessible phenomena, to elicit people's preferences or hidden behaviors and to forecast social or economical events in the medium or the long run?
- How to prune the redundant information that concurrent databases and repetitive records contain? Also, how to screen the statistically valid from the coarse information in excessively loaded databases created for purposes that are alien to statistics?

This is the reason why, in this volume, methodologies for measuring statistical errors and for designing complex questionnaires are picked out. Statistical errors refer to both sampling and non-sampling errors. The paper by Giovanni D' Alessio and Giuseppe Ilardi examines methods for measuring the effects of non-response errors ("unit non-response") and of some response errors ("uncorrelated measurement errors" and "biases from underreporting") by taking advantage of the experience from the Bank of Italy's Survey on Household Income and Wealth. The Authors suggest, too, how to overcome the error effects on estimates through various techniques and models and by means of auxiliary information.

Questionnaire design and question wording instructions, as strategies for preventing data collection errors, are dealt with extensively in the volume. Cristina Davino and Rosaria Romano discuss the case of multi-item scales that are appropriate for the measurement of subjective data, focussing on how individual propensities in the use of scales can be interpreted, in particular within strata of respondents. Luigi Fabbris covers the presentation of batteries of interrelated items for scoring or ranking sets of choice or preference items. Five types of item presentations in questionnaires are crossed with potential estimation models and computer-assisted data collection modes. Caterina Arcidiacono and Immacolata Di Napoli present and apply the so-called Cantril scale that self-anchors the extremes of a scale and is useful in psychological research. Simona Balbi and Nicole Triunfo deal with the age-old problem of closed- and open-ended questions, in the perspective of statistical analysis of data. The Authors tackle, in particular, the problem of transforming textual data, i.e. data that are collected in natural language, into data that can be processed with multivariate statistical methods.

This volume is a consequence of the stimulating debate that animated the workshop "Thinking about methodology and applications of surveys" that took place at the University of Macerata (Italy) in September 2010. The titles of the papers presented in the volume have been proposed to the Authors having in mind the relevant issues of a homogeneous field of study that is usually covered by undirected articles. In each paper, a survey of the scientific literature is discussed and remarks and innovative solutions to face today's survey problems are suggested. All papers aim at balancing formal rigour with simplicity of the presentation style so as to address the book both to practitioners involved in applied survey research and to academics interested in scientific development of surveys.

The editors of the volume like to highlight that all papers have been referred by at least two external experts in the topical field. The editors wish to thank the Authors and also the Referees for their invaluable contribution to the quality of the papers. The external referees involved were: Giorgio Alleva (University of Rome "La Sapienza", Italy), Gianni Betti (University of Siena, Italy), Silvia Biffignandi (University of Bergamo, Italy), Sergio Bolasco (University of Rome "La Sapienza", Italy), Marisa Civardi (Bicocca University in Milan, Italy), Daniela Cocchi (University of Bologna, Italy), Giuseppe Giordano (University of Salerno, Italy), Michael Greenacre (Universitat Pompeu Fabra, Barcelona, Spain), Filomena Maggino (University of Florence, Italy), Alberto Marradi (University of Florence, Italy), Giovanna Nicolini (University of Milan, Italy), Maria Francesca Romano (Scuola Superiore Sant'Anna, Italy).

The editors are open to any contribution of readers who would wish to comment on papers or to propose the Authors other ideas.

Cristina Davino
Luigi Fabbris

Contents

Part I Introduction to Statistical Surveys

Part II Questionnaire Design

Part III Sampling Design and Error Estimation

Part IV Data Integration

Part I
Introduction to Statistical Surveys

Surveys: Critical Points, Challenges and Need for Development

Luigi Biggeri

Abstract The aim of this paper is to review some of the critical issues which need further insight in survey methodology and practice. The focus is on the specific aspects of the survey process grouping the main issues in the following three areas: (i) mode of collection of data and construct of questionnaire; (ii) sampling strategy, design and estimation, to reply to the demands of the users and integration of data; (iii) data dissemination and standardisation. For every issue survey data producers have to fight and ask for development.

1 Introduction

Statistical survey methodology and applications—from the data collection to the dissemination of the results—have improved greatly during the last decades but the discussions in this field are still to the fore in international and national statistical research and in statistical scientific congresses. For example, three scientific events in this field were organized recently in Italy: in 2007, a Satellite conference of the International Association of Survey Statisticians (IASS) on "Small Area Estimation" was held in Pisa (SAE2007), in 2009, the "First Italian Conference on Survey Methodology" took place in Siena (ITACOSM09) and in 2010 a Workshop on "Statistical Surveys: thinking about methodology and applications" was organized in Macerata (SS2010). However, various issues, concerning both the methodology and practice of statistical surveys, still require discussion and improvement; in particular, operative solutions for them must be defined.

The aim of this paper is to review some of the critical issues which need development on the producer side. The presentation of the paper is organized as follows.

L. Biggeri (✉)
Department of Statistics, University of Florence,
Florence, Italy
e-mail: biggeri@ds.unifi.it

C. Davino and L. Fabbris (eds.), *Survey Data Collection and Integration*,
DOI: 10.1007/978-3-642-21308-3_1, © Springer-Verlag Berlin Heidelberg 2013

Section 2 presents some frameworks regarding the relationship between users and producers of survey data and the life cycle of a survey - from the design, data producer and quality assessment perspectives.. The frameworks will allow us to focus on the specific aspects of the survey process that still present critical issues, which will be discussed in Sect. 3 together with the challenges and the need for development of research topics in this field. Some concluding remarks are presented in Sect. 4.

2 Frameworks to Specify and Organize the Presentation of the Critical Issues of Statistical Surveys

Statistical surveys are not the only method for collecting information about the population (i.e., there are administrative record systems, qualitative investigations, observation of the behaviour of persons, randomised experiments, etc.). However, we focus on survey methodology and practice seeking to identify principles about the design, collection, processing and analysis of survey data that are linked to the cost and quality of survey estimates.

The production processes and ‘quality’ of survey data can be defined within two main frameworks, labelled: (i) *Total Quality Management*, and (ii) *Total Survey Error paradigm*. These frameworks are useful for identifying the areas of the critical issues that may need development.

2.1 The Total Quality Management Approach

Since the 1990s, scholars have debated the possibility of referring to the Total Quality Management (TQM) approach for producing statistical information and improving its quality (Groves 1989; Groves and Tortora 1991; U.S. Bureau of Labor Statistics 1994). What we would like to stress from the outset is the importance of discussing the improvement of the quality of statistical survey results, not only as regards checking and evaluating possible sampling and non-sampling errors, but also as an approach to their construction and computation, including continuous checking and revision of the design of their production process to satisfy the *users’ needs*.
Therefore, in general, and in order to decide (by systematic evaluation) which kind of improvements in the statistical measures are necessary and feasible, it is necessary to follow a Total Quality Measurement Model which, at least in part, may be similar to an experimental design in Taguchi’s approach, in order to evaluate the importance of the different methods which are used in the survey and may affect its results.

A simplified framework of the TQM referred to the production of statistical measures is reported in the following Fig. 1.

It is evident from the framework that the starting point for organising a survey and producing statistical measures is the information requirements of the users.

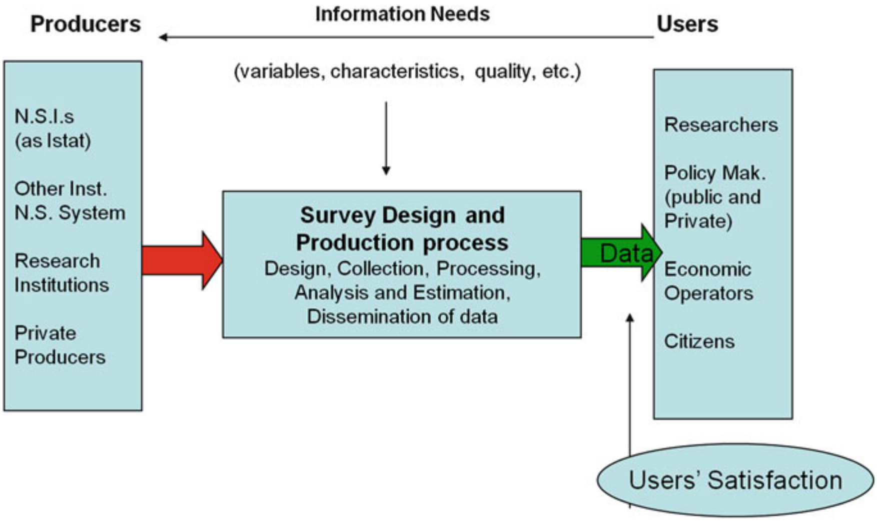

Fig. 1 Total quality management framework for a statistical survey

Three important characteristics of the current production of statistical measures are: (i) the organisation of statistical production must be *user oriented*, looking for the optimisation of the users' satisfaction; (ii) the *emphasis* is placed *on the quality* of all statistical information produced, which is given increasing attention to the check of users' satisfaction; (iii) the resources available for statistical production are scarce, and in any case it is very important to take into account the cost of the survey (*cost constraints*).

The information needs must be specified in terms of variables of interest with their characteristics and, above all, in terms of the quality characteristics of the survey result (of the survey statistics (y)).

With regard to the inputs or process variables (factors), it can be said, in general, that some of them are controllable factors, while other inputs can be considered uncontrollable (or noise) factors; i.e. environmental factors. For an example of application of the framework for the calculation of price indices, see Fig. 2.

In theory, a quality-cost model or cost-error model should serve as a mechanism for evaluating the performance of a statistical measure, insuring good statistical measures while balancing cost and costumer satisfaction. The model could be used to evaluate the cost of alternative designs and to balance each alternative contribution against the analytical goals of the programme. This means that the field focuses on improving quality within cost constraints or, alternatively, reducing costs for a given level of quality.