# Statistical Matching

## Theory and Practice

**Marcello D'Orazio, Marco Di Zio and Mauro Scanu**
*ISTAT – Istituto Nazionale di Statistica, Rome, Italy*

John Wiley & Sons, Ltd

# Statistical Matching

WILEY SERIES IN SURVEY METHODOLOGY
Established in part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

A complete list of the titles in this series appears at the end of this volume.

# Statistical Matching

## Theory and Practice

**Marcello D'Orazio, Marco Di Zio and Mauro Scanu**
*ISTAT – Istituto Nazionale di Statistica, Rome, Italy*

John Wiley & Sons, Ltd

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com



***Other Wiley Editorial Offices***

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

***Library of Congress Cataloging-in-Publication Data***

D'Orazio, Marcello.
Statistical matching : theory and practice / Marcello D'Orazio, Marco Di Zio, and Mauro Scanu.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470-02353-2 (acid-free paper)
ISBN-10: 0-470-02353-8 (acid-free paper)
1. Statistical matching. I. Di Zio, Marco. II. Scanu, Mauro. III. Title.
QA276.6.D67 2006
519.5′2–dc22

2006040184

***British Library Cataloguing in Publication Data***

A catalogue record for this book is available from the British Library

ISBN-13: 978-0-470-02353-2 (HB)
ISBN-10: 0-470-02353-8 (HB)

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by TJ International, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

# Contents

# Preface

Statistical matching is a relatively new area of research which has been receiving increasing attention in response to the flood of data which are now available. It has the practical objective of drawing information piecewise from different independent sample surveys.

The origins of statistical matching can be traced back to the mid-1960s, when a comprehensive data set with information on socio-demographic variables, income and tax returns by family was created by matching the 1966 Tax File and the 1967 Survey of Economic Opportunities; see Okner (1972). Interest in procedures for producing information from distinct sample surveys rose in the following years, although not without controversy. Is it possible to draw joint information on two variables never jointly observed but distinctly available in two independent sample surveys? Are standard statistical techniques able to solve this problem? As a matter of fact, there are two opposite aspects: the practical aspect that aims to produce a large amount of information rapidly and at low cost, and the theoretical aspect that needs to assess whether this production process is justifiable. This book is positioned at the boundary of these two aspects.

Chapters 1–4 are the methodological core of the book. Details of the mathematical-statistical framework of the statistical matching problem are given, together with examples. One of the objectives of this book is to give a complete, formalized treatment of the statistical matching procedures which have been defined or applied hitherto. More precisely, the data sets will always be samples generated by appropriate models or populations (archives and other nonstatistical sources will not be considered). When dealing with sample surveys, the different statistical matching approaches can be justified according to different paradigms. Most (but not all) of the book will rely on a likelihood based inference. The nonparametric case will also be addressed in some detail throughout the book. Other approaches, based on the Bayesian paradigm or on model assisted approaches for finite populations, will be also highlighted. By comparing and contrasting the various statistical matching procedures we hope to produce a synthesis that justifies their use.

Chapters 5–7 are more related to the practical aspects of statistically matching two files. An experience of the construction of a social accounting matrix (Coli *et al.*, 2005) is described in detail, in order to illustrate the peculiarities of the different phases of statistical matching, and the effect of the use of statistical matching techniques without a preliminary analysis of all the aspects.

Finally, sophisticated methods for statistical matching inevitably require the use of computers. The Appendix details some algorithms written in the R language. (the codes are also available on the following webpage: http://www.wiley.com/go/matching).

This book is intended for researchers in the national statistical institutes, and for applied statisticians who face (perhaps for the first time) the problem of statistical matching and could benefit from a structured summary of results in the relevant literature. Readers should possess a background that includes maximum likelihood methods as well as basic concepts in regression analysis and the analysis of contingency tables (some reminders are given in the Appendix). At the same time, we hope the book will also be of interest to methodological researchers. There are many aspects of statistical matching still in need of further exploration.

We are indebted to all those who encouraged us to work on this problem. We particularly thank Pier Luigi Conti, Francesca Tartamella and Barbara Vantaggi for their helpful suggestions and for careful reading of some parts of this book.

The views expressed in this book are those of the authors and do not necessarily reflect the policy of ISTAT.

Marcello, Marco, Mauro
Roma

# 1

# The Statistical Matching Problem

## 1.1 Introduction

Nowadays, decision making requires as much rich and timely information as possible. This can be obtained by carrying out appropriate surveys. However, there are constraints that make this approach difficult or inappropriate.

(i) It takes an appreciable amount of time to plan and execute a new survey. Timeliness, one of the most important requirements for statistical information, risks being compromised.

(ii) A new survey demands funds. The total cost of a survey is an inevitable constraint.

(iii) The need for information may require the analysis of a large number of variables. In other words, the survey should be characterized by a very long questionnaire. It is well established that the longer the questionnaire, the lower the quality of the responses and the higher the frequency of missing responses.

(iv) Additional surveys increase the response burden, affecting data quality, especially in terms of total nonresponse.

A practical solution is to exploit as much as possible all the information already available in different data sources, i.e. to carryout a statistical integration of information already collected. This book deals with one of these data integration procedures: *statistical matching*. Statistical matching (also called data fusion

*Statistical Matching: Theory and Practice* M. D'Orazio, M. Di Zio and M. Scanu

or synthetical matching) aims to integrate two (or more) data sets characterized by the fact that:

(a) the different data sets contain information on (i) a set of common variables and (ii) variables that are not jointly observed;

(b) the units observed in the data sets are different (disjoint sets of units).

**Remark 1.1** Sometimes there is terminological confusion about different procedures that aim to integrate two or more data sources. For instance, Paass (1985) uses the term 'record linkage' to describe the state of the art of statistical matching procedures. Nowadays record linkage refers to an integration procedure that is substantially different from the statistical matching problem in terms of both (a) and (b). First of all, the sets of units of the two (or more) files are at least partially overlapping, contradicting requirement (b). Secondly, the common variables can sometimes be misreported, or subject to change (statistical matching procedures have not hitherto dealt with the problem of the quality of the data collected). The lack of stability of the common variables makes it difficult to link those records in the files that refer to the same units. Hence, record linkage procedures are mostly based on appropriate discriminant analysis procedures in order to distinguish between those records that are actually a match and those that refer to distinct units; see Winkler (1995) and references therein.

A different set of procedures is also called statistical matching. This is characterized by the fact that the two files are completely overlapping, in the sense that each unit observed in one file is also observed in the other file, contradicting requirement (b). However, the common variables are unable to identify the units. These procedures are well established in the literature (see DeGroot *et al*., 1971; DeGroot and Goel 1976; Goel and Ramalingam 1989) and will not be considered in the rest of this book.

A natural question arises: what is meant by integration? As a matter of fact, integration of two or more sources means the possibility of having joint information on the not jointly observed variables of the different sources. There are two apparently distinct ways to pursue this aim.

- Micro approach – The objective in this case is the construction of a *synthetic* file which is *complete*. The file is complete in the sense that all the variables of interest, although collected in different sources, are contained in it. It is synthetic because it is not a product of direct observation of a set of units in the population of interest, but is obtained by exploiting information in the source files in some appropriate way. We remark that the synthetic nature of data is useful in overcoming the problem of confidentiality in the public use of micro files.

- Macro approach – The source files are used in order to have a direct estimation of the joint distribution function (or of some of its key characteristics,

such as the correlation) of the variables of interest which have not been observed in common.

Actually, statistical matching has mostly been analysed and applied following the micro approach. There are a number of reasons for this fact. Sometimes it is a necessary input of some procedures, such as the application of microsimulation models. In other cases, a synthetic complete data set is preferred simply because it is much easier to analyse than two or more incomplete data sets. Finally, joint information on variables never jointly observed in a unique data set may be of interest to multiple subjects (universities, research centres): the complete synthetic data set becomes the source which satisfies the information needs of these subjects.

On the other hand, when the need is just for a contingency table of variables not jointly observed or a set of correlation coefficients, the macro approach can be used more efficiently without resorting to synthetic files. It will be emphasized throughout this book that the two approaches are not distinct. The micro approach is always a byproduct of an estimation of the joint distribution of all the variables of interest. Sometimes this relation is explicitly stated, while in other cases it is implicitly assumed.

Before analysing statistical matching procedures in detail, it is necessary to define the notation and the statistical/mathematical framework for the statistical matching problem; see Sections 1.2 and 1.3. These details will open up a set of different issues that correspond to the different chapters and sections of this book. The outline of the book is given in Section 1.5.

## 1.2 The Statistical Framework

Throughout the book, we will analyse the problem of statistically matching two independent sample surveys, say $A$ and $B$. We will also assume that these two samples consist of records independently generated from appropriate models. The case of samples drawn from finite populations will be treated separately in Chapter 5.

Let $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ be a random variable with density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$, $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, $\mathbf{z} \in \mathcal{Z}$, and $\mathcal{F} = \{f\}$ be a suitable family of densities.[1] Without loss of generality, let $\mathbf{X} = (X_1, \ldots, X_P)'$, $\mathbf{Y} = \left(Y_1, \ldots, Y_Q\right)'$ and $\mathbf{Z} = (Z_1, \ldots, Z_R)'$ be vectors of random variables (r.v.s) of dimension $P$, $Q$ and $R$, respectively. Assume that $A$ and $B$ are two samples consisting of $n_A$ and $n_B$ independent and identically distributed (i.i.d.) observations generated from $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Furthermore, let the units in $A$ have $\mathbf{Z}$ missing, and the units in $B$ have $\mathbf{Y}$ missing. Let

$$\left(\mathbf{x}_a^A, \mathbf{y}_a^A\right) = \left(x_{a1}^A, \ldots, x_{aP}^A, y_{a1}^A, \ldots, y_{aQ}^A\right),$$

[1] We will use the term 'density' for both absolutely continuous and discrete variables, in the former case with respect to the Lebesgue measure, and in the latter case with respect to the counting measure. Hence, in the discrete case $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ should be interpreted as the probability that $\mathbf{X}$ assumes category $\mathbf{x}$, $\mathbf{Y}$ category $\mathbf{y}$ and $\mathbf{Z}$ category $\mathbf{z}$.