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