

# 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

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.<sup>1</sup> Without loss of generality, let  $\mathbf{X} = (X_1, \dots, X_P)'$ ,  $\mathbf{Y} = (Y_1, \dots, Y_Q)'$  and  $\mathbf{Z} = (Z_1, \dots, Z_R)'$  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

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

---

<sup>1</sup>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  $\mathbf{z}$ .

$a = 1, \dots, n_A$ , be the observed values of the units in sample  $A$ , and

$$(\mathbf{x}_b^B, \mathbf{z}_b^B) = (x_{b1}^B, \dots, x_{bP}^B, z_{b1}^B, \dots, z_{bR}^B),$$

$b = 1, \dots, n_B$ , be the observed values of the units in sample  $B$  (for the sake of simplicity, we will omit the superscripts  $A$  and  $B$  and identify the observed values in the two samples by the sample counters  $a$  and  $b$ , unless otherwise specified). When the objective is to gain information on the joint distribution of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  from the observed samples  $A$  and  $B$ , we are dealing with the statistical matching problem.

Table 1.1 shows typical statistical matching samples  $A$  and  $B$ . These samples can be considered as a unique sample  $A \cup B$  of  $n_A + n_B$  i.i.d. observations from  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  characterized by:

- the presence of missing data, and hence of a missing data generation mechanism;
- the absence of joint information on  $\mathbf{X}$ ,  $\mathbf{Y}$ , and  $\mathbf{Z}$ .

The first point has been the focus of a very large statistical literature (see also Appendix A). The possible characterizations of the missing data generation mechanisms for the statistical matching problem are treated in Section 1.3. It will be seen that standard inferential procedures for partially observed samples are also appropriate for the statistical matching problem.

The second issue is actually the essence of the statistical matching problem. Its treatment is the focus throughout this book.

**Remark 1.2** The previous framework for the statistical matching problem has frequently been used (at least implicitly) in practice. However, real statistical matching applications may not fit such a framework. One of the strongest assumptions is that  $A \cup B$  is a unique data set of i.i.d. records from  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ . When, for instance, the two samples are drawn at different times, this assumption may no longer hold.

Without loss of generality, let  $A$  be the most up-to-date sample of size  $n_A$  still from  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  (which is the joint distribution of interest), with  $\mathbf{Z}$  missing. Let  $B$  be a sample independent of  $A$  whose  $n_B$  sample units are i.i.d. from the distribution  $g(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , with  $g$  distinct from  $f$ . It is questionable whether these samples can be statistically matched. Matching can actually be performed when, although the two distributions  $f$  and  $g$  differ, the conditional distribution of  $\mathbf{Z}$  given  $\mathbf{X}$  is the same on both occasions. In this case, appropriate statistical matching procedures have been defined which assign different roles to the two samples  $A$  and  $B$ :  $B$  should lend information on  $\mathbf{Z}$  to the  $A$  file. In the following it will be made clear whenever this alternative framework is under consideration.

Table 1.1 Sample data  $A \cup B$  for the statistical matching problem. The shaded cells correspond to the unobserved variables in samples  $A$  and  $B$ , respectively

| Sample | $Y_1$         | $\dots$ | $Y_q$         | $\dots$ | $Y_Q$         | $X_1$         | $\dots$ | $X_p$         | $\dots$ | $X_P$         | $Z_1$         | $\dots$ | $Z_r$         | $\dots$ | $Z_R$         |
|--------|---------------|---------|---------------|---------|---------------|---------------|---------|---------------|---------|---------------|---------------|---------|---------------|---------|---------------|
| $A$    | $y_{11}^A$    | $\dots$ | $y_{1q}^A$    | $\dots$ | $y_{1Q}^A$    | $x_{11}^A$    | $\dots$ | $x_{1p}^A$    | $\dots$ | $x_{1P}^A$    |               |         |               |         |               |
|        | $\dots$       | $\dots$ | $\dots$       | $\dots$ | $\dots$       | $\dots$       | $\dots$ | $x_{1p}^A$    | $\dots$ | $x_{1P}^A$    |               |         |               |         |               |
|        | $y_{a1}^A$    | $\dots$ | $y_{aq}^A$    | $\dots$ | $y_{aQ}^A$    | $x_{a1}^A$    | $\dots$ | $x_{ap}^A$    | $\dots$ | $x_{aP}^A$    |               |         |               |         |               |
|        | $\dots$       | $\dots$ | $\dots$       | $\dots$ | $\dots$       | $\dots$       | $\dots$ | $x_{n_A p}^A$ | $\dots$ | $x_{n_A P}^A$ |               |         |               |         |               |
|        | $y_{n_A 1}^A$ | $\dots$ | $y_{n_A q}^A$ | $\dots$ | $y_{n_A Q}^A$ | $x_{n_A 1}^A$ | $\dots$ | $x_{n_A p}^A$ | $\dots$ | $x_{n_A P}^A$ |               |         |               |         |               |
| $B$    |               |         |               |         |               | $x_{11}^B$    | $\dots$ | $x_{1p}^B$    | $\dots$ | $x_{1P}^B$    | $z_{11}^B$    | $\dots$ | $z_{1r}^B$    | $\dots$ | $z_{1R}^B$    |
|        |               |         |               |         |               | $\dots$       | $\dots$ | $x_{bp}^B$    | $\dots$ | $x_{bP}^B$    | $\dots$       | $\dots$ | $z_{br}^B$    | $\dots$ | $z_{bR}^B$    |
|        |               |         |               |         |               | $x_{n_B 1}^B$ | $\dots$ | $x_{n_B p}^B$ | $\dots$ | $x_{n_B P}^B$ | $z_{n_B 1}^B$ | $\dots$ | $z_{n_B r}^B$ | $\dots$ | $z_{n_B R}^B$ |
|        |               |         |               |         |               | $\dots$       | $\dots$ | $x_{n_B p}^B$ | $\dots$ | $x_{n_B P}^B$ | $\dots$       | $\dots$ | $z_{n_B r}^B$ | $\dots$ | $z_{n_B R}^B$ |
|        |               |         |               |         |               | $x_{n_B 1}^B$ | $\dots$ | $x_{n_B p}^B$ | $\dots$ | $x_{n_B P}^B$ | $z_{n_B 1}^B$ | $\dots$ | $z_{n_B r}^B$ | $\dots$ | $z_{n_B R}^B$ |

### 1.3 The Missing Data Mechanism in the Statistical Matching Problem

Before going into the details of the statistical matching procedures, let us describe the overall sample  $A \cup B$ . As already described in Section 1.2, it is a sample of  $n_A + n_B$  units from  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  with  $\mathbf{Z}$  missing in  $A$  and  $\mathbf{Y}$  missing in  $B$ . Hence, the statistical matching problem can be regarded as a problem of analysis of a partially observed data set. Generally speaking, when missing items are present, it is necessary to take into account a set of additional r.v.s  $\mathbf{R} = (\mathbf{R}_x, \mathbf{R}_y, \mathbf{R}_z)$ , where  $\mathbf{R}_x$ ,  $\mathbf{R}_y$  and  $\mathbf{R}_z$  are respectively random vectors of dimension  $P$ ,  $Q$  and  $R$ :

$$\mathbf{R}_x = (R_{X_1}, \dots, R_{X_P})',$$

$$\mathbf{R}_y = (R_{Y_1}, \dots, R_{Y_Q})',$$

$$\mathbf{R}_z = (R_{Z_1}, \dots, R_{Z_R})'.$$

The indicator r.v.  $R_{X_j}$  shows when  $X_j$  has been observed ( $R_{X_j} = 1$ ) or not ( $R_{X_j} = 0$ ),  $j = 1, \dots, P$ . Similar definitions hold for the random vectors  $\mathbf{R}_y$  and  $\mathbf{R}_z$ . Appropriate inferences when missing items are present should consider a model that takes into account the variables of interest  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  and the missing data mechanism  $\mathbf{R}$ . Particularly important is the relationship among these variables, defined by the conditional distribution of  $\mathbf{R}$  given the variables of interest:  $h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z})$ . Rubin (1976) defines three different missing data models, which are generally assumed by the analyst: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR); see Appendix A. Indeed, the statistical matching problem has a particular property: missingness is induced by the sampling design. When  $A$  and  $B$  are jointly considered as a unique data set of  $n_A + n_B$  independent units generated from the same distribution  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , with  $\mathbf{Z}$  missing in  $A$  and  $\mathbf{Y}$  missing in  $B$ , i.e. for the statistical matching problem, the missing data mechanism is MCAR. A missing data mechanism is MCAR when  $\mathbf{R}$  is independent of either the observed and the unobserved r.v.s  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$ . Consequently,

$$h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z | \mathbf{x}, \mathbf{y}, \mathbf{z}) = h(\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z). \quad (1.1)$$

In order to show this assertion, it is enough to consider that  $\mathbf{R}$  is independent of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , i.e. equation (1.1), or, equivalently for the symmetry of the concept of independence between r.v.s, that the conditional distribution of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  given  $\mathbf{R}$ , say  $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z)$  does not depend on  $\mathbf{R}$ :

$$\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z}),$$

for every  $\mathbf{x} \in \mathcal{X}$ ,  $\mathbf{y} \in \mathcal{Y}$ ,  $\mathbf{z} \in \mathcal{Z}$ .

As a matter of fact, the statistical matching problem is characterized by just two patterns of  $\mathbf{R}$ :

- $\mathbf{R} = (\mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R)$  for the units in  $A$  and
- $\mathbf{R} = (\mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R)$  for the units in  $B$ ,

where  $\mathbf{1}_j$  and  $\mathbf{0}_j$  are two  $j$ -dimensional vectors of ones and zeros, respectively. Due to the i.i.d. assumption of the generation of the  $n_A + n_B$  values for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , we have that

$$\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R) = \phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R) = f(\mathbf{x}, \mathbf{y}, \mathbf{z}) \quad (1.2)$$

for every  $\mathbf{x} \in \mathcal{X}$ ,  $\mathbf{y} \in \mathcal{Y}$ ,  $\mathbf{z} \in \mathcal{Z}$ , where  $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R)$  is the distribution which generates the records in sample  $A$  and  $\phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R)$  is the distribution which generates the records in sample  $B$ . In other words, the missing data mechanism is independent of both observed and missing values of the variables under study, which is the definition of the MCAR mechanism. This fact allows the possibility of making inference on the overall joint distribution of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  without considering (i.e. *ignoring*) the random indicators  $\mathbf{R}$ . Additionally, inferences can be based on the observed sampling distribution. This is obtained by marginalizing the overall distribution  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  with respect to the unobserved variables. As a consequence, the observed sampling distribution for the  $n_A + n_B$  units is easily computed:

$$\prod_{a=1}^{n_A} f_{\mathbf{XY}}(\mathbf{x}_a, \mathbf{y}_a) \prod_{b=1}^{n_B} f_{\mathbf{XZ}}(\mathbf{x}_b, \mathbf{z}_b). \quad (1.3)$$

The observed sampling distribution (1.3) is the reference distribution for this book, as it is for most papers on statistical matching; see, for instance, Rässler (2002, pg. 78). The following remark underlines which alternatives can be considered, what missing data generation mechanism refers to them, and their feasibility.

**Remark 1.3** Remark 1.2 states that  $A$  and  $B$  cannot always be considered as generated from an identical distribution. In this case, equation (1.1) no longer holds and the missing data mechanism in  $A \cup B$  cannot be assumed MCAR. In the notation of Remark 1.2, the distributions of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  given the patterns of missing data are:

$$\begin{aligned} \phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{1}_Q, \mathbf{0}_R) &= f(\mathbf{x}, \mathbf{y}, \mathbf{z}), \\ \phi(\mathbf{x}, \mathbf{y}, \mathbf{z} | \mathbf{1}_P, \mathbf{0}_Q, \mathbf{1}_R) &= g(\mathbf{x}, \mathbf{y}, \mathbf{z}), \end{aligned}$$

for every  $\mathbf{x} \in \mathcal{X}$ ,  $\mathbf{y} \in \mathcal{Y}$ ,  $\mathbf{z} \in \mathcal{Z}$ . This situation can be formalized via the so-called pattern mixture models (Little, 1993): if the two samples are analysed as a unique sample of  $n_A + n_B$  units, the corresponding generating model is a mixture of the two distributions  $f$  and  $g$ . Little warns that this approach usually leads to unidentified models, and shows which restrictions that tie unidentified parameters with the identified ones should be used. In general, as already underlined in Remark 1.2, the interest is not in the mixture of the two distributions, but only in the most up-to-date one,  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  (an exception will be illustrated in Remark 6.1). For this reason, these models will not be considered any more. The framework illustrated in Remark 1.2 will just consider  $B$  as a donor of information on  $\mathbf{Z}$ , when possible.

## 1.4 Accuracy of a Statistical Matching Procedure

Sections 1.2 and 1.3 have described the input of the statistical matching problem: a partially observed data set with the absence of joint information on the variables of interest and some basic assumptions on the data generating model. This section deals with the output. As declared in Section 1.1, the statistical matching problem may be addressed using either the micro or macro approach. These approaches can be adopted by using many different statistical procedures, i.e. different transformations of the available (observed) data. Are there any guidelines as to the choice of procedure? In other words, how is it possible to assess the accuracy of a statistical matching procedure?

It must be remarked that it is not easy to draw definitive conclusions. Papers that deal explicitly with this problem are few in number, among them Barr and Turner (1990); see also D’Orazio *et al.* (2002) and references therein. A number of different issues should be taken into account.

- (a) What assumptions can be reasonably considered for the joint model  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ ?
- (b) What estimator for  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is preferable, if any, under the model assumed in (a)?
- (c) What method of generating appropriate values for the missing variables can be used under the model chosen in (a) and according to the estimator chosen in (b)?

As a matter of fact, (a) is a very general question related to the data generation process, (b) is related to the macro approach, and (c) to the micro approach. They are interrelated in the sense that an apparently reasonable answer to a question is not reasonable if the previous questions are unanswered. Actually, there is yet another question that should be considered when a synthetic file is distributed and inferential methods are applied to it.

- (d) What inferential procedure can be used on the synthetic data set?

The combination of (a) and (b) for the macro approach, (a), (b) and (c) for the micro approach, and (a), (b), (c), and (d) for the analysis of the result of the micro approach gives an overall sketch of the accuracy of the corresponding statistical matching result. A general measure that amalgamates all these aspects has not been yet defined. It can only be assessed via appropriate Monte Carlo experiments in a simulated framework.

Let us investigate each of the accuracy issues (a)–(d) in more detail.

### 1.4.1 Model assumptions

Table 1.1 shows that the statistical matching problem is characterized by a very annoying situation: there is no observation where all the variables of interest are

jointly recorded. A consequence is that, of all the possible statistical models for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , only a few are actually identifiable for  $A \cup B$ . In other words,  $A \cup B$  does not contain enough information for the estimation of parameters such as the correlation matrix or the contingency table of  $(\mathbf{Y}, \mathbf{Z})$ . Furthermore, for the same reason, it is not possible to test on  $A \cup B$  which model is appropriate for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . There are different possibilities.

- Further information (e.g. previous experience or an *ad hoc* survey) justifies the use of an identifiable model for  $A \cup B$ .
- Further information (e.g. previous experience or an *ad hoc* survey) is used together with  $A \cup B$  in order to make other models also identifiable.
- No assumptions are made on the  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  model. This problem is studied as a problem characterized by uncertainty on some of the model properties.

The first two assumptions are able to produce a unique point estimate of the parameters. For the third choice, which is a conservative one, a set rather than a point estimate of the inestimable parameters, such as the correlation matrix of  $(\mathbf{Y}, \mathbf{Z})$ , will be the output. The features of this set of estimates describe uncertainty for that parameter.

The first two choices are assumptions that should be well justified by additional sources of information. If these assumptions are wrong, no matter what sophisticated inferential machinery is used, the results of the macro and, hence, of the micro approaches will reflect the assumption and not the real underlying model. Also in these cases, evaluation of uncertainty is a precious source of information. In fact, reliability of conclusions based on one of the first two choices can be based on the evaluation of their uncertainty when no assumptions are considered. For instance, if a correlation coefficient for the never jointly observed variables  $\mathbf{Y}$  and  $\mathbf{Z}$  is estimated under a specific identifiable model for  $A \cup B$  or with the help of further auxiliary information, an indication of the reliability of these estimates is given by the width of the uncertainty set: the smaller it is, the higher is the reliability of the estimates with respect to model misspecification.

### 1.4.2 Accuracy of the estimator

Let us assume that a model for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  has been firmly established. When the approach is macro, accuracy of a statistical matching procedure means accuracy of the estimator of the distribution function  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ . In this case, appropriate measures such as the mean square error (MSE) or, accounting for its components, the bias and variance are well known in both the parametric and nonparametric case.

In a parametric framework, minimization of the MSE of each parameter estimator can (almost) be obtained, at least for large data sets and under minimal regularity conditions, when maximum likelihood (ML) estimators are used. More precisely, the consistency property of ML estimators is claimed in most of the results of this book. It must be emphasized that the ML approach given the overall set  $A \cup B$

has an additional property in this case: every parameter estimate is coherent with the other estimates. Sometimes a partially observed data set may suggest distinct estimators for each parameter of the joint distribution that are not coherent. It will be seen that this issue is fundamental in statistical matching, given that it deals with the partially observed data set of Table 1.1.

In a nonparametric framework, consistency of the results is also one of the most important aspects to consider. Consistency of estimators is a very important characterization for the statistical matching problem. In fact, it ensures that, for large samples, estimates are very close to the true but unknown distribution  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ . In the next subsection it will be seen that this aspect is relevant also to the micro approach.

### 1.4.3 Representativeness of the synthetic file

This aspect is the most commonly investigated issue for assessing the accuracy of a statistical matching procedure. Generally speaking, four large categories of accuracy evaluation procedures can be defined (Rässler, 2002), from the most difficult goal to the simplest:

- (i) Synthetic records should coincide with the true (but unobserved) values.
- (ii) The joint distribution of all variables is reflected in the statistically matched file.
- (iii) The correlation structure of the variables is preserved.
- (iv) The marginal and joint distributions of the variables in the source files are preserved in the matched file.

The first point is the most ambitious and difficult requirement to fulfil. It can be achieved when logical or mathematical rules determining a single value for each single unit are available. However, when using statistical rules, it is not as important to reproduce the exact value as it is the joint distribution  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , which contains all the relevant statistical information.

The third and fourth points do not ensure that the final synthetic data set is appropriate for any kind of inferences for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , contradicting one of the main characteristics that a synthetic data set should possess. For instance, the fourth point ensures only reasonable inferences for the distributions of  $(\mathbf{X}, \mathbf{Y})$  and  $(\mathbf{X}, \mathbf{Z})$ .

When the second goal is fulfilled, the synthetic data set can be considered as a sample generated from the joint distribution of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ . Hence, the synthetic data set is *representative* of  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , and can be used as a general purpose sample in order to infer its characteristics.

Any discrepancy between the real data generating model and the underlying model of the synthetic complete data set is called *matching noise*; see Paass (1985).

Focusing on the second point, under identifiable models or with the help of additional information (Section 1.4.1), the relevant question is whether the data

synthetically generated via the estimated distribution  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  are affected by the matching noise or not. It is not always a simple matter. As claimed in Section 1.4.2, when the available data sets are large and the macro approach is used with a consistent estimator of  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , it is possible to define micro approaches with a reduced matching noise. Note that a good estimate of  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is a necessary but not a sufficient condition to ensure that the matching noise is as low as possible. In fact, the generation of the synthetic data should be also done appropriately.

#### 1.4.4 Accuracy of estimators applied on the synthetic data set

This is a critical issue for the micro approach. If the synthetic data set can be considered as a sample generated according to  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$  (or approximately so), it is appropriate to use estimators that would be applied in complete data cases. Hence, the objective of reducing the matching noise (Section 1.4.3) is fundamental.

In fact, estimators preserve their inferential properties (e.g. unbiasedness, consistency) with respect to the model that has generated the synthetic data. When the matching noise is large, these results are a misleading indication as to the true model  $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ .

As a matter of fact, this last problem resembles that of Section 1.4.1. In Section 1.4.1 there was a model misspecification problem. Now the problem is that the data generating model of the synthetic data set differs from the data generating model of the observed data set. In both cases the result is similar: inferences are related to models that differ from the target one.

### 1.5 Outline of the Book

This book aims to explore the statistical matching problem and its possible solutions. This task will be addressed by considering features of its input (Sections 1.2 and 1.3) and, more importantly, of its output (Section 1.4).

One of the key issues is model assumption. As remarked in Section 1.4.1, a first set of techniques refer to the case where the overall model family  $\mathcal{F}$  is identifiable for  $A \cup B$ . A natural identifiable model is one that assumes the independence of  $\mathbf{Y}$  and  $\mathbf{Z}$  given  $\mathbf{X}$ . This assumption is usually called the *conditional independence assumption* (CIA). Chapter 2 is devoted to the description and analysis of the different statistical matching approaches under the CIA.

The set of identifiable models for  $A \cup B$  is rather narrow, and may be inappropriate for the phenomena under study. In order to overcome this problem, further auxiliary information beyond just  $A \cup B$  is needed. This *auxiliary information* may be either in parametric form, i.e. knowledge of the values of some of the parameters of the model for  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ , or as an additional data sample  $C$ . The use of auxiliary information in the statistical matching process is described in Chapter 3.

Both Chapters 2 and 3 will consider the following aspects:

- (i) macro and micro approaches;

- (ii) parametric and nonparametric definition of the set of possible distribution functions  $\mathcal{F}$ ;
- (iii) the possibility of departures from the i.i.d. case (as in Remark 1.2).

As claimed in Section 1.4.1, a very important issue deals with the situation where no model assumptions are hypothesized. In this case, it is possible to study the uncertainty associated to the parameters due to lack of sample information. Given the importance of this topic, it is described in considerable detail in Chapter 4.

The framework developed in Section 1.2 is not the most appropriate for samples drawn from finite populations according to complex survey designs, unless ignorability of the sample design is claimed; see, for example, Gelman *et al.* (2004, Chapter 7). Despite the amount of data sets of this kind, only few methodological results for statistical matching are available. A general review of these approaches and the link with the corresponding results under the framework of Section 1.2 is given in Chapter 5.

Generally speaking, statistical integration of different sources is strictly connected to the integration of the data production processes. Actually, statistical integration of sources would be particularly successful when applied to sources already standardized in terms of definitions and concepts. Unfortunately, this is not always true. Some considerations on the preliminary operations needed for statistically matching two samples are reported in Chapter 6.

Finally, Chapter 7 presents some statistical matching applications. A particular statistical matching application is described in some detail in order to make clear all the tasks that should be considered when matching two real data sets. Furthermore, this example allows the comparison of the results of different statistical matching procedures.

All the original codes used for simulations and experiments, developed in the R environment (R Development Core Team, 2004), are reported in Appendix E in order to enable the reader to make practical use of the techniques discussed in the book. The same codes can also be downloaded on the site <http://www.wiley.com/go/matching>.