Matching information from two independent informative samples

Daniela Marella*, Danny Pfeffermann** [1]

*Dipartimento di Scienze della Formazione, Università "Roma Tre"

**Central Bureau of Statistics and Hebrew University of Jerusalem, Israel;

University of Southampton, UK.

Abstract

Abstract: Statistical matching deals with the problem of how to combine information collected in different samples taken from the same target population, but on partly different survey variables. The purpose of this paper is to analyze the statistical matching problem under informative sampling designs, when applying the sample likelihood approach. First, a conditional independence assumption is made, which allows to define an identifiable population model under which the conditions guaranteeing the identifiability and estimability of the sample likelihood are investigated. Next, the uncertainty in statistical matching under informative sampling designs is discussed, with particular attention to the three-variate normal case. A simulation experiment illustrating the theoretical results is performed.

Key words: conditional independence, informative sampling, matching uncertainty, sample distribution, sample likelihood.

1. Introduction

The statistical matching problem consists of combining information collected from different samples drawn from the same population, with only partial overlap between the two samples. Formally, the problem can be described as follows. Let A and B be two independent samples of sizes n_A and n_B respectively, selected from a population of

[1] email: daniela.marella@uniroma3.it, D.Pfeffermann@soton.ac.uk
Corresponding author: Daniela Marella - Dipartimento di Scienze della Formazione, Università "Roma Tre" - daniela.marella@uniroma3.it

independent records $\{(x_i, y_i, z_i)\}$, generated from some joint density $f(x, y, z; \theta)$, governed by a vector parameter θ. The problem of statistical matching is that (X, Y, Z) are not jointly observed in the two samples. Specifically, suppose that only (X, Y) are observed for the units in sample A, and only (X, Z) are observed for the units in sample B. See, e.g., [17] and [6]. We assume that the samples A and B have no units in common, which will generally be the case if the population is sufficiently large and the selection probabilities are small.

The goal of statistical matching is to reconstruct a matched (fused) data set in which each record includes measurements on (X, Y, Z), which users may treat as a "completely" observed data set from a single source. Due to the absence of joint observations of Z and Y for given X, the distribution $f(x, y, z; \theta)$ is generally not identifiable. In order to overcome this problem, two approaches have been considered in the literature. The first approach assumes conditional independence between Y and Z given X (hereafter CIA); see, e.g., [11]. The CIA has a very important role in statistical matching, the reason being that under the CIA, the distribution of (X, Y, Z) is identifiable and directly estimable from the information provided by the two samples. Appropriateness of the CIA is discussed in several papers. We cite, among others, [22] and [19]. The second approach assumes the existence of external auxiliary information regarding the statistical relationship between Y and Z; see, e.g., [23]. However, commonly, neither approach is applicable. The CIA is rarely met in practice and relevant external auxiliary information is not often available. The lack of joint observations on the variables of interest implies uncertainty about the model holding for (X, Y, Z). In other words, the sample information provided by A and B is not sufficient to enable to distinguish among plausible models for (X, Y, Z), resulting in *identification problems*. In a parametric setting, the consequence of the identification problem is that only ranges of plausible values of the missing records, obtained from models fitted to the available sample information can be defined. Intervals defined by these ranges are known in the literature as *uncertainty intervals.* References tackling the problem of assessing the statistical matching uncertainty in the context of independent and identically observations ($i.i.d$) are [8], [20], [10] and [17]. Uncertainty in statistical matching in a nonparametric setting under the $i.i.d$ assumption is considered in [1], [2] and [3].

In practice, the $i.i.d$ assumption is itself questionable, particularly when dealing with sample survey data. The sample selection in survey sampling involves complex sampling designs based on different levels of clustering and differential inclusion probabilities, which could be correlated with the survey variables of interest. This can violate the independence assumption and result in different distributions of the observed data from the distribution holding in the population from which the samples are drawn. See Section 3. Statistical matching in complex sample surveys is studied in [20], [18] and [4]. [20] proposes to compute new sampling weights for all the units in $A \cup B$. However, this approach is seldom applied since it requires to assess the inclusion probabilities of the units in one sample under the sampling design of the other sample. [18] proposes to calibrate the sampling weights in A and B such that the new weights, when applied to the measured x-values in the two samples, reproduce the known (estimated) population totals of X. Next, the author estimates the joint distribution of categorical variables Y and Z under the CIA. [4] deals with the statistical matching problem for complex sample surveys non-parametrically. The authors propose to estimate the distribution function of variables which are not jointly observed based on an iterative proportional fitting algorithm, and show how to evaluate its reliability.

The aim of the present paper is to analyze the statistical matching problem for the case where the sampling processes used to select the samples A and B are informative for the target variables of interest. As already mentioned, official survey data are usually collected from samples drawn by probability sampling. When the inclusion probabilities are related to the value of the target outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. This, quite common phenomenon, is known in the survey sampling literature as informative sampling. In this case, conventional analysis, which ignores the informative sampling effects may yield large bias and erroneous inference, as illustrated, for example, in the book edited by [24]. See [14] for discussion of the notion of informative sampling and review of methods to deal with this problem. Returning to statistical matching, knowledge of the sampling designs underlying the selection of the samples A and B and accounting for them, is crucial for successful matching. This is true even under the simplified CIA framework.

The paper is organized as follows. In Section 2 we summarize briefly the parametric solution to the matching problem under the CIA, for the case where the sampling process is noninformative. Section 3 considers the case of informative sampling under the CIA and defines the corresponding sample likelihood for the statistical matching problem. The use of the sample likelihood enables estimating the corresponding population distributions and to impute the missing values. The conditions under which the sample models are identifiable and estimable from the information provided by the samples A and B are investigated in Section 4. Section 5 analyses the case of a three-variate normal distribution, investigating the effects of different informative sampling designs on the population model. In Section 6 the CIA assumption is relaxed, and the uncertainty in statistical matching under informative sampling is restudied, illustrated in Section 7, where we again restrict to the three-variate normal case. Section 8 contains results of simulation experiments used to illustrate the theoretical results. We conclude with a brief summary and suggestions for further research in Section 9. All proofs and additional results are deferred to the Supplementary Material.

2. Statistical matching under CIA and noninformative sampling

Suppose that the population values $D_p = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}$ are independent realizations from some joint probability density function (*pdf*) $f(x, y, z; \theta)$, indexed by a vector parameter θ. When the CIA holds, the population model $f_p(x, y, z; \theta)$ can be factorized as

$$f_p(x_i, y_i, z_i; \theta) = f_p(x_i; \theta_X) f_p(y_i|x_i; \theta_{Y|X}) f_p(z_i|x_i; \theta_{Z|X}), \tag{2.1}$$

where the (vector) parameters θ_X, $\theta_{Y|X}$ and $\theta_{Z|X}$ governing the corresponding three distributions, are assumed to be distinct. Under noninformative sampling, such that the population model holds also for the sample data, the model (2.1) is identifiable and directly estimable from the two distinct samples A and B. Furthermore, when the distributions in (2.1) are parametric and the parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$ are distinct, it is possible to resort to maximum likelihood estimation (MLE), the MLE of $\theta = (\theta_X, \theta_{Y|X}, \theta_{Z|X})$ can be obtained by considering separately the corresponding three likelihoods. Once the parametric models have been estimated, a matched data set can be obtained by imputing the missing values in the combined file $A \cup B$ from the distributions of the corresponding missing target variables, given the observed variables.

3. The sample likelihood under CIA and informative sampling

Let V_p^A, V_p^B be sets of population values of design variables used for selecting the samples A and B, respectively. The design variables often contain strata and cluster indicators, and/or variables of measures of size, used for probability proportional to size sampling. The design values V_p^A, V_p^B are known to the persons selecting the samples, but generally not to analysts analysing the data. Some or all of the variables (X, Y, Z) may be included among the design variables, but at least Y and Z are only known to the analyst of the sample data. We assume that the population data, D_p and the values of V_p^A, V_p^B are realizations of random processes, implying that the first order sample inclusion probabilities $\pi_i^A = Pr(i \in A)$, $\pi_i^B = Pr(i \in B)$ may be viewed as random as well. The sample inclusion probabilities contain invaluable information about the relationship between the distribution of the sample data and the distribution in the population from which the samples are taken. Denote by $w_i^A = 1/\pi_i^A$, $w_i^B = 1/\pi_i^B$ the (basic) sampling weights for the two samples. The sample data may be viewed therefore as the outcome of two random processes: the first process generates the values for the N population units, while the second process selects a sample from the finite population according to the underlying sampling design. As it is often the case in a secondary analysis, the data available to the analyst consist of only the two samples $A = \{(x_a, y_a, w_a^A)\}$ and $B = \{(x_b, z_b, w_b^B)\}$ of sizes n_A and n_B, respectively. Let I_a^A be the sample A indicator; $I_a^A = 1$ if population unit $a \in A$, $I_a^A = 0$ otherwise. If Z was observed in A, then following [12], the marginal sample *pdf* of (x_a, y_a, z_a) for $a \in A$ is defined as

$$\begin{aligned} f_A(x_a, y_a, z_a) &= \frac{P(I_a^A = 1|x_a, y_a, z_a)}{P(I_a^A = 1)} f_p(x_a, y_a, z_a) \\ &= f_A(x_a) f_A(y_a|x_a) f_A(z_a|x_a, y_a), \end{aligned} \tag{3.1}$$

and under independence between observations corresponding to different sampling units, (see Remark 1 below), the corresponding sample likelihood based on the sample A (without parameter notation, see below) is,

$$L^A = \prod_{a=1}^{n_A} f_A(x_a, y_a, z_a) = \prod_{a=1}^{n_A} f_A(x_a, y_a) f_A(z_a|x_a, y_a). \tag{3.2}$$

However, since the variable Z is not observed in A, the sample likelihood of θ based on the observed data in A is obtained by integrating the missing data out of the complete sample likelihood (3.2). The *observed sample likelihood* of A is thus given by,

$$L^A_{Obs}(\theta_X, \theta_{Y|X}, \gamma^A) = \int \prod_{a=1}^{n_A} f_A(x_a, y_a, z_a) dz_a = \prod_{a=1}^{n_A} f_A(x_a, y_a; \theta_X, \theta_{Y|X}, \gamma^A) \tag{3.3}$$

where γ^A represents any additional parameters defining the sample distribution, resulting from the sample process. See Equations (3.6) and (3.7) below.

Remark 1. *[12] establish general conditions under which for independent observations under the population model, the sample measurements are asymptotically independent under the sample model, when increasing the population size but holding the sample size fixed. This permits approximating the sample likelihood by the product of the sample* pdfs *over the corresponding sample observations. In Section 4 we discuss the identifiability and estimability of the sample* pdf.

Similarly, the *observed sample likelihood* of B is,

$$L^B_{Obs}(\theta_X, \theta_{Z|X}, \gamma^B) = \int \prod_{b=1}^{n_B} f_B(x_b, y_b, z_b) dy_b = \prod_{b=1}^{n_B} f_B(x_b, z_b; \theta_X, \theta_{Z|X}, \gamma^B). \tag{3.4}$$

Hence, the *sample likelihood* of the sample $A \cup B$ is,

$$\begin{aligned} L^{A\cup B}_{Obs}(\theta, \gamma^A, \gamma^B) &= \prod_{a=1}^{n_A} f_A(x_a, y_a; \theta_X, \theta_{Y|X}, \gamma^A) \prod_{b=1}^{n_B} f_B(x_b, z_b; \theta_X, \theta_{Z|X}, \gamma^B) \\ &= \prod_{a=1}^{n_A} f_A(y_a|x_a; \theta_{Y|X}, \gamma^A) \prod_{b=1}^{n_B} f_B(z_b|x_b; \theta_{Z|X}, \gamma^B) \\ & \prod_{a=1}^{n_A} f_A(x_a; \theta_X, \gamma^A) \prod_{b=1}^{n_B} f_B(x_b; \theta_X, \gamma^B), \end{aligned} \tag{3.5}$$

where the parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$, governing the population *pdf*s, are orthogonal because of the conditional independence of Y and Z given X. Notice that (3.5) has a similar structure to the observed data likelihood under noninformative sampling, as defined in [17], the big difference being that the population models $f_p(x_a, y_a)$, $f_p(x_b, z_b)$

are replaced in (3.5) by the sample models $f_A(x_a, y_a)$, $f_B(x_b, z_b)$. The maximum likelihood estimates are obtained by maximizing separately the corresponding three sample likelihoods in the right hand side of (3.5), where the estimator of θ_X is obtained by maximizing $\prod_{a=1}^{n_A} f_A(x_a) \prod_{b=1}^{n_B} f_B(x_b)$, thus utilizing the data in both samples. Following [12] and [13], the sample *pdf*s featuring in (3.5) can be expressed alternatively as,

$$\begin{aligned} f_A(x_a; \theta_X, \gamma^A) &= \frac{P(I_a^A = 1|x_a; \gamma^A)}{P(I_a^A = 1; \theta_X, \gamma^A)} f_p(x_a; \theta_X) = \frac{E_p(\pi_a^A|x_a; \gamma^A)}{E_p(\pi_a^A; \theta_X, \gamma^A)} f_p(x_a; \theta_X) \\ &= \frac{E_A(w_a^A; \theta_X, \gamma^A)}{E_A(w_a^A|x_a; \gamma^A)} f_p(x_a; \theta_X) \end{aligned} \tag{3.6}$$

where $f_p(x_a; \theta_X)$ represents the corresponding population *pdf.* Equation (3.6) uses the relationship between the population *pdf* and the sample *pdf.* Notice that the expectations in the left- and right hand side of this relationship refer to different distributions. The relationship $E_p(\pi_a^A|x_a; \gamma^a) = 1/E_A(w_a^A|x_a; \gamma^A)$ has been established in [13], where E_p, E_A denote expectations under the population and sample distributions, respectively. When $P(I_a^A = 1|x_a; \gamma^A) = P(I_a^A = 1; \theta_X, \gamma^A)$ for each x_a, the population and sample models are the same and the sampling design may be ignored for inference on the parameters θ_X. The conditional marginal sample *pdf* $f_A(y_a|x_a; \theta_{Y|X}, \gamma^A)$ is defined as the conditional *pdf* of $y_a|x_a$ given that unit a is in the sample A. Similarly to (3.6),

$$f_A(y_a|x_a; \theta_{Y|X}, \gamma^A) = \frac{E_A(w_a^A|x_a; \theta_{Y|X}, \gamma^A) f_p(y_a|x_a; \theta_{Y|X})}{E_A(w_a^A|x_a, y_a; \gamma^A)}. \tag{3.7}$$

Similar expressions to (3.6) and (3.7) are obtained for the sample *pdf*s $f_B(x_b)$ and $f_B(z_b|x_b)$ operating in the sample B. The sample distributions are functions of the corresponding population *pdf*s and the respective conditional expectations of the sampling weights. The parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$ governing the three population models can be estimated from the corresponding sample data, by MLE. Furthermore, the expectations displayed in the sample *pdf*s can be estimated from the observed data, using classical model fitting procedures. Fixing the unknown parameters featuring in these expectations at their estimated values allows to maximize the resulting likelihoods only with respect to the parameters indexing the population *pdf*s, thus simplifying and stabilizing the maximization process. For example, the expectation $E_A(w_a^A|x_a, y_a; \gamma^A)$ in (3.7) can be estimated by regressing w_a^A against (x_a, y_a), using the observations $\{(x_a, y_a, w_a^A), a \in A\}$. See, e.g., [14] and [15] for examples of regression models that can be used for this purpose, depending on the problem at hand. Alternatively, the expectation $E_A(w_a^A|x_a, y_a; \gamma^A)$ can be estimated non-parametrically, as in [7]. Once the parameters governing the population model have been estimated, a fused dataset $D_f = \{(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i), i = 1, \ldots, \tilde{n}\}$, where $\tilde{n}$ is the desired sample size, (possibly $\tilde{n} = N$, the population size), can be constructed, such that each record includes measurements of (X, Y, Z), which users may treat as a "completely" observed dataset, with a similar distribution to the population distribution. Formally, one may employ the following procedure: (i) Generate $\tilde{n}$ observations $\tilde{x}_i$ from $f_p(x_i; \widehat{\theta}_X)$; (ii) For each $i = 1, \ldots, \tilde{n}$ draw at random a value $\tilde{y}_i$ from $f_p(y_i|x_i; \widehat{\theta}_{Y|X})$ (iii) For each $i = 1, \ldots, \tilde{n}$ draw at random a value $\tilde{z}_i$ from $f_p(z_i|x_i; \widehat{\theta}_{Z|X})$.

Under general likelihood theory, the population model parameter estimators are consistent, guaranteeing that for sufficiently large samples A and B, the synthetic dataset D_f defines a genuine sample from the population *pdf* $f_p(x_i, y_i, z_i; \theta)$.

Remark 2. *Unlike in the case of noninformative sampling, it is generally not correct to only impute the missing* z*-values for the sample* A*, and the missing* y*-values for the sample* B*, and consider the union of the two samples as the final matched data set of size* $n_A + n_B$*. The reason for this is that the sample* $A \cup B$ *is informative and does not represent the population distribution adequately. Thus, although the* z*-values can be generated from the population model* $f_p(z_i|x_i; \widehat{\theta}_{Z|X})$*, the observed* x*- and* y*-values may not represent the corresponding population* x*- and* y*-values adequately.*

4. Identifiability and estimability of the sample model

As outlined in Section 3, we base the inference on the unknown model parameters on the sample likelihood, which requires that the corresponding sample *pdf*s $f_A(x_a, y_a)$ and $f_B(x_b, z_b)$ are identifiable and estimable from the available sample information. The sample model $f_A(x_a, y_a)$ is identifiable, if no different pairs $[f_p^1(x_a, y_a), P^1(I_a^A = 1|x_a, y_a)]$, $[f_p^2(x_a, y_a), P^2(I_a^A = 1|x_a, y_a)]$ exist, which induce the same sample *pdf* $f_A(x_a, y_a)$. See [16] and references therein for conditions guaranteeing the identifiability of the sample model. Similar conditions are required for the identifiability of the sample model $f_B(x_b, z_b)$. By estimability we mean that the population and sampling parameters appearing in the sample models can be estimated from the available data when using the sample likelihood for estimation. As stated in Proposition 1 below, the estimability of the vector parameter $\theta = (\theta_X, \theta_{Y|X}, \theta_{Z|X})$ based on the likelihood (3.5) depends on both the fulfillment of the CIA at the sample level and the properties of the sampling process.

Remark 3. *Even if the CIA holds in the population, it does not necessarily hold in the sample because even conditionally on X, the sampling process could induce dependence between the not jointly observed variables Y and Z. See Example 3 in Section 5.*

Proposition 1. *If the CIA holds for the population values and $P(I_a^A = 1|x_a, y_a, z_a) = P(I_a^A = 1|x_a, y_a)$ for all z_a then,*

1. *The CIA holds in sample A .*

2. *The sample model $f_A(x_a, y_a)$ is estimable from the sample information in A.*

Proof of Proposition 1 is in the Supplementary Material, Section S1.

Remark 4. *Under the conditions of Proposition 1, it is possibile to impute the missing Z values in sample A and obtain complete observations from the sample model $f_A(x_a, y_a, z_a)$, by estimating $f_p(z|x)$ from sample B. However, as already noted in Remark 2, the* *extended sample may be informative and not represent the population distribution adequately.*

It is not correct to conclude from Proposition 1 that if the sampling process in A depends on the z-values, the sampling and the population parameters featuring in $f_A(x_a, y_a)$ can not be estimated. The sampling probabilities π_a^A may depend on many unobserved design variables including Z and yet, by definition of the sample *pdf* $f_A(x_a, y_a)$, one only needs to model the probability $P(I_a^A = 1|x_a, y_a)$. As discussed in Section 3, the latter probability can be estimated from the sample data by regressing the sampling weights against (x_a, y_a). Moreover, the resulting sample model (3.7) can be tested, as discussed and illustrated in [15] and [7]. A similar proposition to Proposition 1 applies to the sample B.

Remark 5. *In the case of noninformative sampling, it is sufficient to assume the CIA for proper inference on the population parameters. However, under informative sampling, one needs also to verify that the sample* pdfs *featuring in the likelihood (3.5) can be estimated and tested on the basis of the available information for the samples A and B.*

5. Informative sampling from a trivariate normal population satisfying the CIA

In this section we consider the case where the population distribution is trivariate normal that satisfies the CIA. In particular, we study the effect of alternative informative sampling designs on the inference process. Suppose that (X, Y, Z) is normal with mean $\mu = (\mu_X, \mu_Y, \mu_Z)$ and variance-covariance (V-C) matrix Σ, such that $\sigma_{YZ} = \sigma_{YX}\sigma_{ZX}/\sigma_X^2$. Under the CIA, the population distribution can be factorized as in (2.1), where the marginal distribution of X is $N(\mu_X, \sigma_X^2)$ and the conditional distribution of Y given $X = x$ is $N(\mu_{Y|x}, \sigma_{Y|X}^2)$, with $\mu_{Y|x} = \beta_0 + \beta_1 x$, $\beta_0 = \mu_Y - \beta_1\mu_X$, $\beta_1 = \sigma_{XY}/\sigma_X^2$, $\sigma_{Y|X}^2 = \sigma_Y^2 - \beta_1^2\sigma_X^2$. Similar expressions hold for the conditional distribution of Z given $X = x$ with parameters $(\alpha_0, \alpha_1, \sigma_{Z|X}^2)$. In Examples 1, 2 and 4 below, (but not in Example

3), the CIA is preserved by the sampling process. Furthermore, the sample model and the population model are in the same family and only differ in some or all the parameters. Yet, since the units in A have missing Z values while the units in B have missing Y values, the reference sample model is the sample *pdf* $f^{sm}_{A\cup B}(x_i, y_i, z_i)$,

$$f^{sm}_{A\cup B}(x_i, y_i, z_i) \quad = \quad f_{A\cup B}(x_i) f_A(y_i|x_i) f_B(z_i|x_i). \tag{5.1}$$

Additionally, in Examples 1, 2 the samples A and B are selected with the same sampling probabilities. Clearly, in real situations the sampling designs giving rise to the two samples might be different as considered in Example 4.

Example 1. *Following [12], suppose that the sample inclusion probabilities in A and B have expectations,*

$$E_p(\pi_i|x_i, y_i, z_i) = \kappa \exp\{\gamma_X x_i + \gamma_Y y_i + \gamma_Z z_i\}, \tag{5.2}$$

where κ guarantees that the expectation is less or equal to one. With regard to Equation (5.2), under probability proportional to size sampling

$$E_p(\pi_i|x_i, y_i, z_i) = N n_A \frac{\exp\{\gamma_X x_i + \gamma_Y y_i + \gamma_Z z_i\}}{N \sum_{j=1}^{N} \exp\{\gamma_X x_j + \gamma_Y y_j + \gamma_Z z_j\}}. \tag{5.3}$$

For large N (the common case), $\sum_{j=1}^{N} \exp\{\gamma_X x_j + \gamma_Y y_j + \gamma_Z z_j\}/N$ can be considered as a constant and so is κ, in which case it cancels out in the derivation of the sample distribution. See [12] and [13] for discussion and many other examples.

Under (5.2), the joint sample pdf *can be factorized as,*

$$f_S(x_i, y_i, z_i) = f_S(x_i) f_S(y_i|x_i) f_S(z_i|x_i), \tag{5.4}$$

where $S = A, B$ denotes the sample and *(a)* $f_S(x_i)$ *is* $N(\mu_X + (\gamma_X + \beta_1\gamma_Y + \alpha_1\gamma_Z)\sigma_X^2, \sigma_X^2)$*;*

(b) $f_S(y_i|x_i)$ *is* $N(\mu_{Y|x_i} + \gamma_Y\sigma_{Y|X}^2, \sigma_{Y|X}^2)$*;*

(c) $f_S(z_i|x_i)$ *is* $N(\mu_{Z|x_i} + \gamma_Z\sigma_{Z|X}^2, \sigma_{Z|X}^2)$*;*

(d) the sample pdf $f_{A\cup B}^{sm}(x_i, y_i, z_i)$ *(5.1), is* $N(\mu_{A\cup B}, \Sigma_{A\cup B})$*, where the sample V-C matrix* $\Sigma_{A\cup B}$ *is the same as for the population distribution,* $\Sigma_{A\cup B} = \Sigma$*,*

thus showing that the CIA holds for the sample. See Supplementary Material, Section S2.

Example 2. *Following [12], suppose that the sample inclusion probabilities in A and B have expectations,*

$$E_p(\pi_i|x_i, y_i, z_i) = \kappa \exp\{\gamma_X x_i + \gamma_Y y_i + \gamma_{2Y} y_i^2\}, \tag{5.5}$$

with $\gamma_{2Y} < 0$*. The joint sample* pdf *can be factorized in this case as in (5.4) where*

(a) $f_S(x_i)$ *is* $N(\mu_X + (\gamma_X + \beta_1\gamma_Y)\sigma_X^2, \sigma_X^2)$*;*

(b) $f_S(y_i|x_i)$ *is* $N\left(\frac{\mu_{Y|x_i}+\gamma_Y\sigma_{Y|X}^2}{1-2\gamma_{2Y}\sigma_{Y|X}^2}, \frac{\sigma_{Y|X}^2}{1-2\gamma_{2Y}\sigma_{Y|X}^2}\right)$*;*

(c) $f_S(z_i|x_i)$ *is* $N(\mu_{Z|x_i}, \sigma_{Z|X}^2)$*;*

(d) the joint sample pdf $f_{A\cup B}^{sm}(x_i, y_i, z_i)$ *(5.1) is* $N(\mu_{A\cup B}, \Sigma_{A\cup B})$*;*

(e) for $\gamma_{2Y} < 0$*, the sample correlation matrix* $\Phi_{A\cup B}$ *is positive semidefinite.*

thus showing that the CIA holds for the sample. Note that by Proposition 1, $f_S(z_i|x_i) = f_p(z_i|x_i)$ *with* $S = A, B$*. See Supplementary Material, Section S3.*

Example 3. *Suppose that the sample inclusion probabilities in A and B have expectations,*

$$E_p(\pi_i|x_i, y_i, z_i) = \kappa \exp\{\gamma_X x_i + \gamma_{YZ} y_i z_i\}. \tag{5.6}$$

In this case, the CIA does not hold for the sample, and the sample pdf *is no longer trivariate normal. See Supplementary Material, Section S4.*

Example 4. *Suppose that the sampling inclusion probabilities in A and B have expectation given by (5.2) and (5.5), respectively;*

$$\begin{aligned} E_p(\pi_i^A|x_i,y_i,z_i) &= \kappa_A \exp\{\gamma_X^A x_i + \gamma_Y^A y_i + \gamma_Z^A z_i\}, \\ E_p(\pi_i^B|x_i,y_i,z_i) &= \kappa_B \exp\{\gamma_X^B x_i + \gamma_Y^B y_i + \gamma_{2Y}^B y_i^2\} \end{aligned} \tag{5.7}$$

(different expectations in the two samples). The joint sample pdf $f^{sm}_{A\cup B}(x_i,y_i,z_i)$ *(5.1) is* $N(\mu_{A\cup B},\Sigma)$*. See Supplementary Material, Section S5.*

6. Uncertainty analysis under informative sampling

So far, we assumed that the joint population distribution satisfies the CIA. However, the CIA may not hold in practice and the lack of joint measurements on the variables (X,Y,Z) disallows to distinguish between different plausible distributions that possibly hold for them. Let H_p be the set of population *pdf*s $f_p(x,y,z;\theta)$ having bivariate marginal *pdf*s $f_p(x,y;\theta_{XY})$ and $f_p(x,z;\theta_{XZ})$, which we assume to be known except for their underlying parameters.

$$\begin{aligned} H_p[f_p(x,y),f_p(x,z)] = \Big\{ f_p(x,y,z) : \int f_p(x,y,z)dz = f_p(x,y), \\ \int f_p(x,y,z)dy = f_p(x,z) \Big\}. \end{aligned} \tag{6.1}$$

Each distribution in H_p is viewed as a plausible joint distribution of the variables (X,Y,Z). Following [4], we refer to each such distribution as a *matching distribution.* The larger the class H_p, the more uncertain is the model for (X,Y,Z). At the sample level, the matching distribution should be chosen from the class defined by (6.1), but with bivariate *pdf*s estimated on the basis of the sample data. Yet, the parameters θ_{YZ} (part of the vector θ indexing the *pdf* $f_p(x, y, z; \theta)$) cannot be estimated from the samples A and B, implying that instead of point estimates, one can only construct a set of plausible values for θ_{YZ}, which are consistent with the estimates of θ_{XY} and θ_{XZ}. Each plausible estimate of θ_{YZ} defines a plausible model in the class H_p. Note, however, that it is not possible to prioritize one model over another, given the available sample data. As noted before, we assume that the form of the population distribution $f_p(x, y, z; \theta)$ is known so that the uncertainty is with regard to the parameters θ_{YZ}. Suppose that a value θ^*_{YZ} is chosen from the set of plausible values of θ_{YZ} and $f_p(x, y, z; \theta^*_{YZ}) = f^*_p(x, y, z) \in H_p$ is chosen as the matching distribution of (X, Y, Z), whereas the true population *pdf* is $f_p(x, y, z) \in H_p$. The discrepancy between $f^*_p(x, y, z)$ and $f_p(x, y, z)$ is the *matching error.* The smaller the matching error, the closer is the *pdf* $f^*_p(x, y, z)$ to the true population *pdf* $f_p(x, y, z)$. Let $\widehat{f^*_p}(x, y, z)$ be an estimate of $f^*_p(x, y, z)$ for a chosen θ^*_{YZ}. By [4], setting $t = (x, y, z)$ and $dt = dxdydz$, the total estimation error can be decomposed as,

$$\begin{aligned}\int_{I\!R^k} |\widehat{f^*_p}(t) - f_p(t)|dt \;\leqslant\; & \int_{I\!R^k} |\widehat{f^*_p}(t) - f^*_p(t)|dt \\ & + \int_{I\!R^k} |f^*_p(t) - f_p(t)|dt \qquad (6.2)\end{aligned}$$

where $k = dim(x, y, z)$. The first term in the right hand side of (6.2) is the *sampling error* due to estimation of the other parameters in θ, which can be estimated consistently from the samples A and B. For given θ^*_{YZ}, the consistency of $\widehat{f^*_p}(x, y, z)$ guarantees that this term becomes negligible for large sample sizes n_A and n_B. The second term represents the *population model uncertainty* or identification uncertainty, see in [9]. Obviously, the most favorable case, which occurs for instance if the CIA holds, is when the class H_p consists of a single *pdf.* In this case, the population model for (X, Y, Z) is identifiable and directly estimable from the available sample information.

Ignoring the informativness of the sampling designs introduces an implicit assumption that the population model holds for the sample data, in which case the class (6.1) is defined as the set of plausible distributions for (X, Y, Z) having bivariate marginal *pdf*s $f_A(x, y)$, $f_B(x, z)$, that is, the class $H_p[f_A(x, y), f_B(x, z)]$. Let $f^*_{AB}(x, y, z)$ be the population *pdf* chosen from such a class and $\widehat{f^*_{AB}}(x, y, z)$ be an estimate of $f^*_{AB}(x, y, z)$. Similarly to (6.2), the total estimation error can be decomposed now as,

$$\begin{aligned}\int_{I\!R^k} |\widehat{f^*_{AB}}(t) - f_p(t)|dt \;\leqslant\; & \int_{I\!R^k} |\widehat{f^*_{AB}}(t) - f^*_{AB}(t)|dt + \int_{I\!R^k} |f^*_{AB}(t) - f^*_p(t)|dt \\ & + \int_{I\!R^k} |f^*_p(t) - f_p(t)|dt. \qquad (6.3)\end{aligned}$$

Analogously to (6.2), the first term in the right hand side of (6.3) is the *sampling error* due to estimation. The second term represents now the matching error due to the informativness of sampling designs in A and B. The last term is the *population model uncertainty* as defined in (6.2).

Remark 6. *In [4] the notion of total estimation error is dealt in terms of cumulative distribution functions. In section 6 the total estimation error is defined in terms of density functions, so as to be consistent with what is done in previous sections where the effect of informative sampling on the sample* pdf *has been evaluated. Clearly, the use of density functions does not affect the interpretation of the results.*

7. Informative sampling from a trivariate normal population not satisfying the CIA

In this section we analyze the effect of ignoring informative sampling designs for the case of a trivariate normal population distribution, but without imposing the CIA. Consider first the case of noninformative sampling. For this case, [8] shows that the only non-estimable unrestricted parameter is $\rho_{YZ|X}$, the correlation between Y and Z given X. On the other hand, the unconditional correlation ρ_{YZ} is not unrestricted, because of the presence of the common variable X. Assuming that the V-C matrix Σ is positive semidefinite, the correlation ρ_{YZ} must be in the interval,

$$[\tau_p, \nu_p] = [\rho_{YX}\rho_{ZX} - A(\rho_{YX}, \rho_{ZX}), \rho_{YX}\rho_{ZX} + A(\rho_{YX}, \rho_{ZX})] \tag{7.1}$$

where $A(\rho_{YX}, \rho_{ZX}) = \sqrt{(1-\rho_{YX}^2)(1-\rho_{ZX}^2)}$. All the values in the interval (7.1) are equally plausible for ρ_{YZ}. Under the CIA, the parameter ρ_{YZ} is located at the midpoint of the interval (7.1), see [10]. The class H_p of plausible population *pdf*s for (X, Y, Z) given by (6.1) is the set of three-variate normal distributions $f_p(x, y, z)$, with ρ_{YZ} in the interval (7.1). All the distributions in H_p have the same mean vector μ, and the V-C matrices only differ in the entry of σ_{YZ} (or ρ_{YZ} in the correlation matrix). Each value of ρ_{YZ} in the interval (7.1) is associated with one, and only one, plausible population *pdf* for (X, Y, Z). Consequently, choosing a distribution from the class H_p as a candidate for the actual joint *pdf*, is equivalent to choosing a value for ρ_{YZ} in the interval (7.1). The larger the class H_p, the more uncertain is the model for (X, Y, Z). A simple and natural measure of the population model uncertainty, under complete marginal knowledge, is therefore $(\nu_p - \tau_p)$, (Eq. 7.1). The wider the interval, the more uncertain is ρ_{YZ} and hence, the greater is the uncertainty regarding $f_p(x, y, z)$. For further details on the uncertainty measures in parametric settings, see [17] and [6].

Next, consider the case of informative sampling. In this case, a set of maximum likelihood estimates for ρ_{YZ}, (the likelihood ridge, as defined in [6]), can be evaluated. After computing the ML estimates of ρ_{YX} and ρ_{ZX} as described in Section 3 $(\widehat{\rho}_{YX}, \widehat{\rho}_{ZX})$, the likelihood ridge for ρ_{YZ} in the informative case can be evaluated by substituting $(\widehat{\rho}_{YX}, \widehat{\rho}_{ZX})$ in the interval (7.1).

Remark 7. *When information regarding a statistical model for* (Y, Z) *or* $(Y, Z|X)$ *is available, some models for* (X, Y, Z) *might be excluded from the class (6.1), and the statistical model for* (X, Y, Z) *becomes less uncertain, see [17], [5] and [3]. Under a parametric setting, such information assumes the form of constraints on the values of* *nonestimable parameters. For example, in the normal case, the information may consist of constraints on the values of the correlation ρ_{YZ} or equivalently, $\rho_{YZ|X}$.*

In what follows we study the effect of ignoring informative sampling designs on the class H_p and the uncertainty measure $(\nu_p - \tau_p)$, under the sampling processes of Examples 1, 2 and 4, for the case of the trivariate normal population distribution. When the CIA does not hold, the reference sample model in the statistical matching context is given by

$$f^{sm}_{A\cup B}(x_i, y_i, z_i) \quad = \quad f_{A\cup B}(x_i) f_A(y_i|x_i) f_B(z_i|x_i, y_i). \tag{7.2}$$

Example 5. *Under the expectations (5.2), the joint sample* pdf *is*

$$f_S(x_i, y_i, z_i) = f_S(x_i) f_S(y_i|x_i) f_S(z_i|x_i, y_i) \tag{7.3}$$

where,

(a) $f_S(x_i)$ *is* $N(\mu_X + (\gamma_X + \gamma_Z \beta_{ZX|Y} + (\gamma_Y + \beta_{ZY|X}\gamma_Z)\beta_1)\sigma^2_X, \sigma^2_X)$;

(b) $f_S(y_i|x_i)$ *is* $N\left(\mu_{Y|x_i} + (\gamma_Y + \gamma_Z \beta_{ZY|X})\sigma^2_{Y|X}, \sigma^2_{Y|X}\right)$;

(c) $f_S(z_i|x_i, y_i)$ *is* $N(\mu_{Z|x_i y_i} + \gamma_Z \sigma^2_{Z|XY}, \sigma^2_{Z|XY})$;

(d) the joint sample pdf $f^{sm}_{A\cup B}(x_i, y_i, z_i)$ *(7.2) is* $N(\mu_{A\cup B}, \Sigma)$ *with* $\mu_{A\cup B}$ *defined in the Supplementary Material, Equation (S6.6);*

and $S = A, B$. *The coefficient* $\beta_{ZX|Y}$ $(\beta_{ZY|X})$ *is the partial regression coefficients of* Z *on* X *given* Y *(the partial regression coefficients of* Z *on* Y *given* X*),* $\sigma^2_{Z|XY}$ *is the residual variance in the regression of* Z *on* X *and* Y*. When* Y *and* Z *are conditionally independent given* X*, we are back to the results in Example 1. Ignoring the informative sampling designs implies in this case the assumption that the sample model coincides with the population model. Since* $\Sigma_{A\cup B} = \Sigma$ $(\Phi_{A\cup B} = \Phi$ *where* $\Phi_{A\cup B}$ *and* Φ *are the* *corresponding sample and population correlation matrices), the range of plausible values for ρ_{YZ}, obtained under the added assumption that the sample correlation matrix is positive semidefinite, remains unchanged. In this case, ignoring the informative sampling designs affects the composition of the class H_p since the mean vector changes from μ to $\mu_{A\cup B}$, but it does not affect the size of the class or the uncertainty measure $(\nu_p - \tau_p)$. See Supplementary Material, Section S6.*

Example 6. *Under the expectations (5.5), we obtain that*

(a) $f_S(x_i)$ *is* $N(\mu_X + (\gamma_X + \gamma_Y\beta_1)\sigma_X^2, \sigma_X^2)$;

(b) $f_S(y_i|x_i)$ *is* $N\left(\frac{\mu_{Y|x_i}+\gamma_Y\sigma^2_{Y|X}}{1-2\gamma_{2Y}\sigma^2_{Y|X}}, \frac{\sigma^2_{Y|X}}{1-2\gamma_{2Y}\sigma^2_{Y|X}}\right)$;

(c) $f_S(z_i|x_i, y_i)$ *is* $N(\mu_{Z|x_i,y_i}, \sigma^2_{Z|XY})$;

(d) the joint sample pdf $f^{sm}_{A\cup B}(x_i, y_i, z_i)$ *(7.2) is* $N(\mu_{A\cup B}, \Sigma_{A\cup B})$*, with parameters defined in the Supplementary Material, Equations (S7.1), (S7.2);*

where $S = A, B$. When Y and Z are conditionally independent given X, we are back to the results of Example 2. In this case, ignoring the informative sampling design affects both the class H_p both the uncertainty measure. See Supplementary Material, Section S7.

Example 7. *Under the expectations (5.7), the sample* pdf $f^{sm}_{A\cup B}(x_i, y_i, z_i)$ *(7.2) is* $N(\mu_{A\cup B}, \Sigma)$*. Ignoring the informative sampling designs affects in this case the composition of the class H_p, but not the uncertainty measure $(\nu_p - \tau_p)$. See Supplementary Material, Section S8.*

8. Simulation study

In order to illustrate the effects of ignoring the sampling process in statistical matching and to assess the performance of the imputation method described in Section 3, we performed a simulation study as described below.

8.1. Simulation set-up

The simulation experiment consists of the following four steps:

Step 1 Generate $N = 2000$ independent population measurements (x_i, y_i, z_i) from the following trivariate normal distribution satisfying the CIA:

1.1 x_i is normal with parameters $\theta_X = (\mu_X, \sigma_X^2)$; $\mu_X = 5$, $\sigma_X^2 = 1$;

1.2 $y_i|x_i$ is normal with parameters $\theta_{Y|X} = (\beta_0 + \beta_1 x_i; \sigma_{Y|X}^2)$; $\beta_0 = 2$, $\beta_1 = 1$, $\sigma_{Y|X}^2 = 2$;

1.3 $z_i|x_i$ is normal with parameters $\theta_{Z|X} = (\alpha_0 + \alpha_1 x_i; \sigma_{Z|X}^2)$; $\alpha_0 = 1$, $\alpha_1 = 0.5$, $\sigma_{Z|X}^2 = 2$.

Under the CIA, $\rho_{YZ} = \rho_{YX}\rho_{ZX} = 0.19$.

Step 2 Draw independently samples A and B of size $n_A = n_B = 400$ from the population generated in Step 1 by Poisson sampling, with selection probabilities

$$\begin{aligned} \pi_i^A &= n_A \frac{\exp(\gamma_X^A x_i + \gamma_Y^A y_i)}{\sum_{i=1}^N \exp(\gamma_X^A x_i + \gamma_Y^A y_i)}; \\ \pi_i^B &= n_B \frac{\exp(\gamma_X^B x_i + \gamma_Z^B z_i)}{\sum_{i=1}^N \exp(\gamma_X^B x_i + \gamma_Z^B z_i)}. \end{aligned} \tag{8.1}$$

We use different vectors $\gamma^A = (\gamma_X^A, \gamma_Y^A)$ and $\gamma^B = (\gamma_X^B, \gamma_Z^B)$, so as to distinguish between informative and non-informative samples. As shown in Example 1, these sampling probabilities preserve the CIA at the sample level. Notice that despite of the relatively large sampling fraction $f = (400/2000)$, for the sampling designs listed in Table 1 below, the percentage of common units in the samples A and B varies between 0.72% and 1.2%.

Step 3 Construct a fused data set of size $\tilde{n} = N = 2000$ in which the variables (X, Y, Z) are jointly observed, as described in Section 3. For this, the population model parameters $\theta_X, \theta_{Y|X}, \theta_{Z|X}$ are estimated in three different ways:

3.1 By ignoring the sample selection effects. Denote by $f_1(x_i, y_i, z_i)$ the estimated population distribution obtained in this case.

3.2 By assuming that the sampling probability coefficients γ^A, γ^B are known and maximizing the sample likelihood (3.5) with respect to $\theta_X, \theta_{Y|X}, \theta_{Z|X}$. Denote by $f_2(x_i, y_i, z_i)$ the estimated population distribution obtained in this case.

3.3 By maximizing the observed sample likelihood (3.5) with respect to $\theta_X, \theta_{Y|X}, \theta_{Z|X}$, but where the expectations $E_A(w_a^A|x_a, y_a; \gamma^A)$ and $E_B(w_b^B|x_b, z_b; \gamma^B)$ appearing in the sample *pdf*s are also estimated. First, the expectation $E_A(w_a^A|x_a, y_a; \gamma^A)$ is estimated by a linear regression of w_a^A on (x_a, y_a). Second, $E_A(w_a^A|x_a; \gamma^A)$ is evaluated as the integral of $E_A(w_a^A|x_a, y_a; \gamma^A)$ with respect to the conditional sample *pdf* $f_A(y_a|x_a)$, and the integral is inserted into the likelihood, with the unknown sampling parameters γ^A set at their estimated values, so that the likelihood is maximized with respect to the population parameters $\theta_{Y|X}$. A similar procedure is applied for estimating $E_B(w_b^B|x_b, z_b; \gamma^B)$ and $\theta_{Z|X}$. Finally, in order to estimate the parameter θ_X, the expectations $E_A(w_a^A; \gamma^A)$ and $E_B(w_b^B; \gamma^B)$ are expressed as the integrals of $E_A(w_a^A|x_a; \gamma^A)$ and $E_B(w_b^B|x_b; \gamma^B)$ with respect to the sample *pdf*s $f_A(x_a)$ and $f_B(x_b)$, with the corresponding sampling parameters set at their estimated parameters. Notice that by (8.1), the samples A and B are informative not only with respect to Y and Z, but also with respect to X. Denote by $f_3(x_i, y_i, z_i)$ the estimated population distribution obtained in this case.

Step 4 Repeat Steps 2 and 3 $M = 500$ times for each choice of the coefficients γ^A, γ^B defining the sample selection probabilities. We generated the population values only once (Step 1), so as to assess the design-based properties of the various estimation procedures.

8.2. Simulation results

We start by studying the effect of ignoring the informative sampling mechanisms used for drawing the samples A and B, when constructing the fused data set in Step 3. This is done by comparing the estimated population distributions $f_1(x_i, y_i, z_i)$ and $f_2(x_i, y_i, z_i)$. The results are shown in Tables 1-2 where the distance between the true marginal population *pdf*s, $f_p(x_i)$, $f_p(y_i)$, $f_p(z_i)$ and the corresponding estimated marginal *pdf*s $f_h(x_i)$, $f_h(y_i)$, $f_h(z_i)$, for $h = 1, 2$, are reported. In Table 3, the distance between the true population *pdf*, $f_p(x_i, y_i, z_i)$, and the estimated *pdf*s $f_h(x_i, y_i, z_i)$, for $h = 1, 2, 3$ is evaluated. As a measure of distance, we use the symmetric metric,

$$KL_{p,h}(f_p, f_h) = 0.5(KL_{ph}(f_p, f_h) + KL_{hp}(f_h, f_p)) \tag{8.2}$$

where $KL_{ph}(f_p, f_h)$ is the Kullback-Leibler divergence between the two estimated *pdf*s f_p and f_h. We computed for each of the 500 samples the metric (8.2) and then averaged the 500 values, which is viewed as the global divergence measure. Clearly, the smaller the average, the closer on average is the estimated population *pdf* f_h to the true population *pdf* f_p, and the better should be the constructed data set in terms of mirroring the true data set. In Table 1 we report the KL-divergences (8.2) between the population *pdf* $f_p(x)$ and the estimated *pdf*s $f_h(x)$, $h = 1, 2$. The predictive *pdf*s $f_h(x)$ are estimated from the sample data in A, B and $A \cup B$. The corresponding KL-divergences are denoted as $KL_{p,h}^{X,A}$, $KL_{p,h}^{X,B}$ and $KL_{p,h}^{X,A\cup B}$. Although in pratice the *pdf* $f_h(x)$ would be estimated from the sample $A \cup B$, in Table 1 we also report the KL-divergences when the predictive *pdf*s $f_h(x)$; $h = 1, 2$ are estimated based only on the data in A or in B. This is done to illustrate how the informativness of the sampling designs acting in A and B combine in defining the sample model $f_{A\cup B}(x_i)$ and consequently, the KL-divergence $KL_{p,h}^{X,A\cup B}$, for $h = 1, 2$.

For $\gamma^A = (0, 0)$ $KL_{p,1}^{X,A}$ coincides with $KL_{p,2}^{X,A}$ since the sampling process acting in A

Table 1: $KL^{X,A}_{p,h}$, $KL^{X,B}_{p,h}$, $KL^{X,A\cup B}_{p,h}$, h=1,2, for different γ^A, γ^B coefficients.

γ^A	γ^B	$KL^{X,A}_{p,1}$	$KL^{X,A}_{p,2}$	$KL^{X,B}_{p,1}$	$KL^{X,B}_{p,2}$	$KL^{X,A\cup B}_{p,1}$	$KL^{X,A\cup B}_{p,2}$
(0.5,0)	(0,0)	0.125	0.002	0.002	0.002	0.031	0.001
(0, 0)	(0.5, 0)	0.002	0.002	0.126	0.002	0.031	0.001
(0, 1)	(0, 0)	0.257	0.021	0.002	0.002	0.056	0.003
(0, 0)	(0, 1)	0.002	0.002	0.091	0.010	0.020	0.003
(0.5, 1)	(0.5, 1)	0.500	0.055	0.288	0.037	0.379	0.043

is not informative. When $\gamma^A \neq (0,0)$, $KL^{X,A}_{p,1}$ is always larger than $KL^{X,A}_{p,2}$. The same is true for $KL^{X,B}_{p,1}$ and $KL^{X,B}_{p,2}$ when $\gamma^B \neq (0,0)$. Thus, ignoring the sample selection process affects negatively the quality of the predictions of X. Finally, since the sampling designs in A and B combine in defining the sample model $f_{A\cup B}(x_i)$, $KL^{X,A\cup B}_{p,1}$ is always between $KL^{X,A}_{p,1}$ and $KL^{X,B}_{p,1}$ and $KL^{X,A\cup B}_{p,1}$ is always larger than $KL^{X,A\cup B}_{p,2}$. Furthermore, the larger the informativness of the sampling processes, the larger is the distance between f_p and f_h in Table 1. Clearly, such a distance depends also on the characteristics of population *pdf*, that is, on the correlation structure between the variables of interest. For example, since $\rho_{YX} = 0.58 > \rho_{ZX} = 0.33$ ignoring the sampling processes with $\gamma^A = (0,1)$ and $\gamma^B = (0,0)$ yields $KL^{X,A\cup B}_{p,1} = 0.056$, which is larger than $KL^{X,A\cup B}_{p,1} = 0.020$ obtained when $\gamma^A = (0,0)$ and $\gamma^B = (0,1)$. In what follows, the predictive *pdf*s $f_h(x)$ are estimated from the sample $A \cup B$. Table 2 shows the KL-divergences, $KL^{Y}_{p,h}$, $KL^{Z}_{p,h}$, between the marginal population *pdf*s $f_p(y)$, $f_p(z)$, and the corresponding estimated sample *pdf*s, $f_h(y)$, $f_h(z)$; $h = 1,2$. As in Table 1, $KL^{Y}_{p,1}$ and $KL^{Z}_{p,1}$ are always larger than $KL^{Y}_{p,2}$ and $KL^{Z}_{p,2}$, respectively. Thus, ignoring the sample selection process affects negatively the quality of the predictions of Y and Z.

Table 2: $KL^{Y}_{p,h}$, $KL^{Z}_{p,h}$, h=1,2, for different γ^A, γ^B coefficients.

γ^A	γ^B	$KL^{Y}_{p,1}$	$KL^{Y}_{p,2}$	$KL^{Z}_{p,1}$	$KL^{Z}_{p,2}$
(0.5,0)	(0,0)	0.011	0.002	0.006	0.002
(0, 0)	(0.5, 0)	0.012	0.002	0.006	0.002
(0, 1)	(0, 0)	0.769	0.089	0.009	0.002
(0, 0)	(0, 1)	0.008	0.002	0.740	0.044
(0.5, 1)	(0.5, 1)	1.081	0.140	0.915	0.068

In order to evaluate the performance of the imputation procedure proposed in Section 3, we computed the KL-divergence $KL_{p,h}^{XYZ}$ between the true population *pdf* $f_p(x_i, y_i, z_i)$ and the estimated predictive models $f_h(x_i, y_i, z_i)$, for $h = 1, 2, 3$. The results are presented in Table 3 for $\gamma^B = (0, 0)$ and different γ^A coefficients. The last column contains the mean sample size $M(n_A)$ over the 500 A samples selected in each case.

Table 3: $KL_{p,h}^{XYZ}$, h=1,2,3 and mean sample size $M(n_A)$ over the 500 samples for different γ^A coefficients.

γ^A	$KL_{p,1}^{XYZ}$	$KL_{p,2}^{XYZ}$	$KL_{p,3}^{XYZ}$	$M(n_A)$
(0.5, 0)	0.462	0.417	0.421	365
(0, 1)	1.272	0.509	0.530	379
(0.5,1)	1.389	0.574	0.612	402

As expected, ignoring the sample selection effects results in large KL measures. What we find encouraging is that the KL measures when estimating all the unknown parameters are not much larger than the corresponding measures when the sampling parameters are taken as known.

Tables 4 and 5 show the means and standard deviations (Sd) of the 500 estimates of the population parameters $\theta_{Y|X} = (\beta_0, \beta_1, \sigma^2_{Y|X})$ under the predictive models $f_h(x_i, y_i, z_i)$; $h = 1, 3$. The tables help assessing the quality of the potential imputations obtained from the use of the two distributions. The means are denoted as $\overline{\widehat{\beta}}_{0,h}$, $\overline{\widehat{\beta}}_{1,h}$, $\overline{\widehat{\sigma}}^2_{Y|X,h}$; $h = 1, 3$.

Table 4: Means of estimates of $(\beta_0, \beta_1, \sigma^2_{Y|X})$ over the 500 samples with different γ^A coefficients. True parameters are $\beta_0 = 2$, $\beta_1 = 1$, $\sigma^2_{Y|X} = 2$.

γ^A	$\overline{\widehat{\beta}}_{0,1}$	$\overline{\widehat{\beta}}_{0,3}$	$\overline{\widehat{\beta}}_{1,1}$	$\overline{\widehat{\beta}}_{1,3}$	$\overline{\widehat{\sigma}}^2_{Y\|X,1}$	$\overline{\widehat{\sigma}}^2_{Y\|X,3}$
(0.5, 0)	1.90	2.05	1.03	0.98	1.95	2.02
(0,1)	5.08	2.10	0.69	0.97	1.29	1.99
(0.5,1)	5.84	2.15	0.55	0.90	1.30	1.95

As in Table 3, predictions based on $f_1(x_i, y_i, z_i)$ which ignores the sample selection effects produce a synthetic data set with distribution which differs from the true underlying population distribution. Consequently, subsequent inferential procedures

Table 5: Standard deviations of estimates of $(\beta_0, \beta_1, \sigma^2_{Y|X})$ over the 500 samples with different γ^A coefficients.

γ^A	$sd(\widehat{\beta}_{0,1})$	$sd(\widehat{\beta}_{0,3})$	$sd(\widehat{\beta}_{1,1})$	$sd(\widehat{\beta}_{1,3})$	$sd(\widehat{\sigma}^2_{Y\|X,1})$	$sd(\widehat{\sigma}^2_{Y\|X,3})$
(0.5,0)	0.32	0.37	0.06	0.06	0.04	0.04
(0, 1)	0.27	0.86	0.04	0.10	0.03	0.24
(0.5, 1)	0.30	1.23	0.05	0.14	0.03	0.36

based on this data set will be subject to bias, even though the estimates obtained by ignoring the sample selection effects have the smallest variances, a well known phenomenon from other studies. For one of the 500 imputed data sets with $\gamma^A = (0.5, 1)$, Figure 1 shows the population *pdf*, the kernel density estimate of the sample *pdf*, and the distribution of the imputed Y-values when using the predictive distribution $f_3(x_i, y_i, z_i)$. The bandwidth selection rule is as proposed in [21]. Similar results are obtained when using the average of the estimated parameters, reported in Table 4. As clearly seen, the sample *pdf* is very different from the population *pdf* due to the use of informative sampling, but the distribution of the imputed population values is close to the true population distribution.

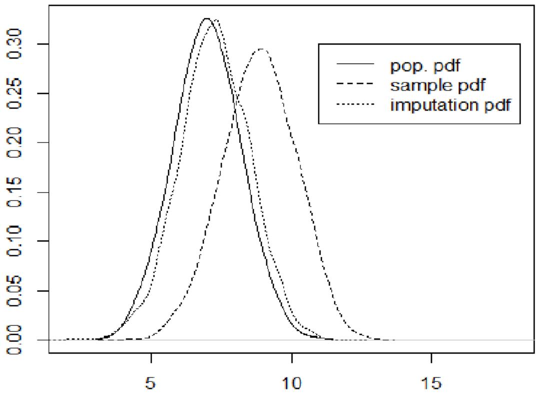

Figure 1: Population *pdf*, Kernel density estimate of the sample *pdf* when ignoring the sample selection effects and distribution of the imputed Y-values.

Finally, Table 6 contains results of estimation of the correlation coefficient between Y and

Z for the case where $\gamma^B = (0,0)$ and alternative sets of sampling coefficients γ^A. In order to evaluate the robustness of the procedure to deviations from the normality assumptions, we consider three different scenarios regarding the true population distribution:

Scenario 1 a three-variate normal distribution as in Step 1, with $\rho^{(1)}_{YZ} = 0.19$;

Scenario 2 x_i has a gamma distribution with shape 3 and scale 1; $y_i|x_i$ and $z_i|x_i$ are normal with parameters as in Step 1, with $\rho^{(2)}_{YZ} = 0.39$;

Scenario 3 x_i has a gamma distribution with shape 3 and scale 1; $y_i|x_i$ is normal as in Step 1; $log(z_i|x_i)$ is normal with parameters $\theta_{Z|X} = (\alpha_0 + \alpha_1 x_i; \sigma^2_{Z|X})$ with $\alpha_0 = 0.1$, $\alpha_1 = 0.4$ and $\sigma^2_{Z|X} = 1$, with $\rho^{(3)}_{YZ} = 0.30$.

For each scenario 500 samples have been drawn from the corresponding population model and then 500 imputed data sets have been generated and used for estimating the correlation. In Table 6, $\overline{\rho}^{(t)}_{YZ,1}$ and $\overline{\rho}^{(t)}_{YZ,3}$, $t = 1,2,3$ represent the means of the correlation estimates from the 500 imputed data sets under the three scenarios, as obtained by use of the imputed values from the estimated predictive distributions $f_1(x_i, y_i, z_i)$ and $f_3(x_i, y_i, z_i)$.

Table 6: Correlation estimates under the three scenarios with different γ^A coefficients. True correlations are $\rho^{(1)}_{YZ} = 0.19$, $\rho^{(2)}_{YZ} = 0.39$, $\rho^{(3)}_{YZ} = 0.30$.

γ^A	$\overline{\rho}^{(1)}_{YZ,1}$	$\overline{\rho}^{(1)}_{YZ,3}$	$\overline{\rho}^{(2)}_{YZ,1}$	$\overline{\rho}^{(2)}_{YZ,3}$	$\overline{\rho}^{(3)}_{YZ,1}$	$\overline{\rho}^{(3)}_{YZ,3}$
(0.5, 0)	0.21	0.19	0.48	0.39	0.49	0.36
(0, 1)	0.18	0.19	0.47	0.39	0.50	0.39
(0.5, 1)	0.16	0.18	0.46	0.38	0.46	0.35

Under the first scenario, the means of the estimated correlation coefficients based on the predictive distributions $f_1(x_i, y_i, z_i)$ and $f_3(x_i, y_i, z_i)$ are similar with small empirical bias. However, under the second and third scenarios, the mean estimates when ignoring the sample selection effects show large empirical bias, where as the estimates obtained when accounting for the sampling effects reveal only small bias.

9. Summary

In this paper we analyzed the statistical matching problem when the two sampling processes used to select the samples A and B are informative. The conditions guaranteeing the identifiability and estimability of the sample likelihood are investigated. When the CIA does not hold, an uncertainty analysis is carried out, showing how ignoring the sampling selection can affect the matching error. Furthermore, the effect of alternative informative sampling designs on the inference process for a trivariate normal population is studied. Finally, a simulation experiment has been performed showing that the magnitude of the error will depend on both the informativeness of the sampling processes and the correlations between the variables of interest. As our paper shows, accounting for the sampling designs effects improves the quality of the matched data file very significantly. The proposed procedure based on maximization of the sample likelihood reveals good performance in terms of the Kullback-Leibler divergence and population parameters estimation and hence in enabling good estimation of the population distribution and imputation of missing observations. Clearly, more theoretical and empirical studies with different population distributions and sampling designs are needed to further ascertain the results of the present paper.

Supplementary Materials The Supplementary Material contains the proof of Proposition 1 and of the results in Examples 1-7.

Acknowledgment The authors are grateful to the referee for very careful reading of the manuscript and thoughtful comments. We dedicate this paper to the memory of Professor Susanne Rässler, who passed away last year while still young and very active in the area of statistical matching.

References

[1] Conti, P.L., Marella, D. and Scanu, M. (2012) Uncertainty Analysis in Statistical Matching. *Journal of Official Statistics*, **28**, pp.1–21.

[2] Conti, P.L., Marella, D. and Scanu, M. (2013) Uncertainty Analysis for Statistical Matching of ordered categorical variables. *Computational Statistics & Data Analysis*, **68**, pp.311–325.

[3] Conti, P.L., Marella, D. and Scanu, M. (2015) How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework. *Communications in Statistics - Theory and Methods*, **1**, 2, pp.967–994.

[4] Conti, P.L., Marella, D. and Scanu, M. (2016) Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, **111**, 516, pp.1715–1725.

[5] D'Orazio, M., Di Zio, M. and Scanu, M. (2006a). Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Official Statistics*, **22**, pp.137–157.

[6] D'Orazio, M., Di Zio, M. and Scanu, M. (2006b). *Statistical Matching: Theory and Practice.* Chichester: Wiley.

[7] Feder, M. and Pfeffermann, D. (2018). Statistical Inference Under Non-ignorable Sampling and Non-response-An Empirical Likelihood Approach. *Paper available from the authors.*

[8] Kadane, J.B., (2001). Some statistical problems in merging data files. *Journal of Official Statistics*, **17**, pp.423-433.

[9] Manski, C. F. (2003). *Partial Identification of Probability Distributions.* New York: Springer-Verlag.

[10] Moriarity, C. and Scheuren, F., (2001). Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*, **17**, pp.407-422.

[11] Okner, B. (1972). Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economic and Social Measurement*, **1**, pp.325-342.

[12] Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distribution of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, pp.1087–1114.

[13] Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā : The Indian Jounal of Statistics* **61**, pp.166–186.

[14] Pfeffermann, D. and Sverchkov, M. (2009). *Inference under Informative Sampling. In: Handbook of Statistics 29B; Sample Surveys: Inference and Analysis. Eds. D. Pfeffermann and C.R. Rao.* North Holland.

[15] Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, **37**, 2, pp.115–136.

[16] Pfeffermann, D., and Landsman, V. (2011). Are private schools really better than public schools? Assessment by methods for observational studies. *Annals of Applied Statistics*, **5**, pp.1726–1751.

[17] Rässler, S.(2002). *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches.* New York: Springer.

[18] Renssen, R. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, **24**, pp.171–183.

[19] Rodgers, W.L. (1984). An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics*, **1**, pp.91–102.

[20] Rubin, D.B. (1986). Statistical Matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economics Statistics*, **4**, pp.87–94.

[21] Sheather, S.J. and Jones, M.C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society Series B*, **53**, pp.683–690.

[22] Sims, C. A. (1972). Comments on: “Constructing a New Data Base From Existing Microdata Sets: the 1966 Merge File”, by B.A. Okner. *Annals of Economic and Social Measurements*, **1**, pp.343–345.

[23] Singh, A.C., Mantel, H., Kinack, M. and Rowe, G., (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Survey Methodology*, **19**, pp.59-79.

[24] Skinner, C.J., Holt, D. and Smith, M.F. (1989). *Analysis of complex surveys.* Wiley.