

Uncertainty in Statistical matching for Complex Sample Surveys

Pier Luigi Conti, Daniela Marella and Mauro Scanu

Abstract An important feature of statistical matching is the estimation of the underlying joint distribution of variables separately available from independent sample surveys. Unless special assumptions are made, the absence of joint information on the variables of interest leads to uncertainty about the data generating model. The aim of this paper is to analyze the uncertainty in statistical matching for complex survey data.

Abstract *Uno degli aspetti di maggior interesse delle tecniche di abbinamento statistico (statistical matching) riguarda la stima della distribuzione congiunta di variabili casuali che sono state osservate distintamente in due indagini campionarie indipendenti per le quali non è possibile effettuare abbinamenti esatti delle unità selezionate. A meno di assunzioni restrittive, l'assenza di osservazioni congiunte delle variabili di interesse produce incertezza sul modello generatore dei dati. L'obiettivo di questo lavoro è l'analisi dell'incertezza che caratterizza il problema dello statistical matching quando i campioni sono generati secondo piani di campionamento complessi.*

Key words: Statistical matching, uncertainty, complex sample surveys

Pier Luigi Conti
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, e-mail: pier-luigi.conti@uniroma1.it

Daniela Marella
Dipartimento di Scienze della Formazione, Università "Roma Tre", e-mail: daniela.marella@uniroma3.it

Mauro Scanu
ISTAT, Italian National Statistical Institute, e-mail: scanu@istat.it

1 Introduction

Statistical matching aims at combining information available in different sample surveys referred to the same target population. Let U_N be a finite population of size N where three characters X, Y, Z are defined. Denote by x_i, y_i, z_i the values taken by X, Y, Z respectively, for unit i ($= 1, 2, \dots, N$) and by X_N, Y_N, Z_N the sequences:

$$X_N = (x_1 \dots, x_N), Y_N = (y_1 \dots, y_N), Z_N = (z_1 \dots, z_N). \quad (1)$$

Let further A and B be two independent complex samples of n_A and n_B records selected from U_N according to the sampling designs p_A and p_B , respectively. More specifically, let $D_{i,A}$ ($D_{i,B}$) be a Bernoulli random variable, such that i is in the sample A (B) whenever $D_{i,A} = 1$ ($D_{i,B} = 1$), whilst i is not in the sample A (B) whenever $D_{i,A} = 0$ ($D_{i,B} = 0$). In particular $\pi_{i,A} = E_{p_A}(D_{i,A})$ and $\pi_{i,B} = E_{p_B}(D_{i,A})$ are the first inclusion probabilities of unit i in sample A and B , respectively.

Assume that the variables (X, Y) are observed in sample A , and that (X, Z) are independently observed in B . The main goal of statistical matching, at a macro level, consists in estimating the joint distribution of (X, Y, Z) . The lack of joint information on the variables of interest is the cause of uncertainty about the model for (X, Y, Z) , leading to the lack of identifiability of the model for (X, Y, Z) . A noticeable exception occurs when Y and Z are independent conditionally on X (CIA, for short).

The uncertainty in statistical matching in a nonparametric setting when A and B are composed by *i.i.d.* observations from (X, Y, Z) has been analyzed in [2], [3]. In [2], the notion of uncertainty in statistical matching is discussed and an overall measure of uncertainty is proposed. Furthermore, the effect on model uncertainty due to the introduction of logical constraints is evaluated. Such constraints include restrictions on the support of the joint distribution of $(Y, Z)|X$. In [3] the uncertainty for order categorical variables is thoroughly investigated.

The sample selection in most surveys involves complex sampling designs based on stratification, different levels of clustering and measures of size. The two main references for statistical matching in complex sample surveys are [5] and [4]. The approach proposed by [5] consists in concatenating the two files A and B by computing new sampling weights relative to an artificial design $p(S)$, where $S = A \cup B$, given by the union of the sampling designs acting in A and B , respectively. Assuming $\pi_{i,A \cap B}$ is negligible, the inclusion probability of a unit i is $\pi_{i,S} = \pi_{i,A} + \pi_{i,B}$. This approach has been seldom used since it requires knowledge of inclusion probabilities of the units in one sample under the sampling design of the other sample. The approach proposed in [4] consists in adapting the actual survey weights of the two distinct samples A and B in order to have homogeneous common distributions on $X, Y|X$ and $Z|X$ through the repeated application of calibration procedures. Estimations based on such A and B samples can be easily defined under the CIA (this can be avoided only through the use of a complete third sample C with joint observations on X, Y, Z under two methods: incomplete-two-way stratification and synthetic-two-way stratification). In this paper we investigate the statistical matching problem by estimates based on A and B distinctly without using any calibration procedures.

2 Uncertainty in statistical matching

2.1 Uncertainty in statistical matching for i.i.d. observations

Let

$$F_N(y, x) = \frac{1}{N} \sum_{i=1}^N I_{(x_i \leq x)} I_{(y_i \leq y)}, \quad G_N(z, x) = \frac{1}{N} \sum_{i=1}^N I_{(x_i \leq x)} I_{(y_i \leq z)}, \quad Q_N(x) = \frac{1}{N} \sum_{i=1}^N I_{(x_i \leq x)}$$

be the joint population distribution functions (p.d.f.s) of (X, Y) , (X, Z) and the marginal p.d.f. of X , respectively, and let $p_N(x)$ be the proportion of population units such that $x_i = x$. Define further the conditional p.d.f.s

$$F_N(y|x) = \frac{F_N(y, x)}{p_N(x)}, \quad G_N(z|x) = \frac{G_N(z, x)}{p_N(x)}, \quad H_N(y, z|x) = \frac{H_N(x, y, z)}{p_N(x)} \quad (2)$$

(note that (2) can be arbitrarily defined whenever $p_N(x) = 0$). Knowledge of the p.d.f.s $F_N(y|x)$, $G_N(z|x)$ does not imply knowledge of $H_N(y, z|x)$. If only the p.d.f.s (2) were known, then one could only say that

$$\max(0, F_N(y|x) + G_N(z|x) - 1) \leq H_N(y, z|x) \leq \min(F_N(y|x), G_N(z|x)). \quad (3)$$

The bounds in (3) are the well-known *Fréchet bounds*. Fréchet bounds (3) can be improved when auxiliary information is available. The kind of constraints we consider in the present paper is $a_x \leq f(y, z) \leq b_x$ given $X = x$, where $f(y, z)$ is a monotone function of y (z) for each z (y). In case of *i.i.d.* observations, such logical constraints were first discussed in [2], and used in [3] in the special case of discrete ordinal variates Y, Z . Under the constraint, the Fréchet bounds (3) are given by $K_{N-}^x(y, z) \leq H_N(y, z|x) \leq K_{N+}^x(y, z)$, where

$$K_{N-}^x(y, z) = \max(0, G_N(z|x) \wedge G_N(\gamma_y(a_x)|x) + F_N(y|x) \wedge F_N(\delta_z(b_x)|x) - 1) \\ K_{N+}^x(y, z) = \min(G_N(z|x), G_N(\gamma_y(a_x)|x), F_N(y|x), F_N(\delta_z(b_x)|x))$$

and $\gamma_y(\cdot)$, $\delta_z(\cdot)$ being the inverse functions of $f(y, z)$ for fixed y and z , respectively. Under the above constraint, and using the same arguments as in [2], a natural measure of uncertainty on $H_N(y, z|x)$ *conditionally on x* is

$$\Delta^x(F_N, G_N) = \frac{1}{N^2 p_N(x)^2} \sum_{i=1}^N \sum_{j=1}^N (K_{N+}^x(y_i, z_j) - K_{N-}^x(y_i, z_j)) I_{(x_i=x)} I_{(x_j=x)} \\ = \int_{R^2} (K_{N+}^x(y, z) - K_{N-}^x(y, z)) d[F_N(y|x) G_N(y|x)] \quad (4)$$

while an *unconditional uncertainty* measure of the (X, Y, Z) joint distribution is

$$\Delta(F_N, G_N) = \sum_x \Delta^x(F_N, G_N) p_N(x). \quad (5)$$

Clearly, the unconditional uncertainty measure (5) is the average of the conditional uncertainty measures (4), w.r.t. the marginal distribution of X .

2.2 Uncertainty in statistical matching for complex survey data

In order to make inference on the uncertainty measures in complex survey data it is necessary to make assumptions on the sampling designs according to which the samples A, B are drawn, as in [1]. The conventional design-based estimator of $F_N(y|x)$ and $G_N(z|x)$ are given by

$$\hat{F}_H(y|x) = \frac{\sum_{i \in U} \frac{D_{i,A}}{\pi_{i,A}} I_{(y_i \leq y)} I_{(x_i = x)}}{\sum_{i \in U} \frac{D_{i,A}}{\pi_{i,A}} I_{(x_i = x)}} \quad (6)$$

$$\hat{G}_H(y|x) = \frac{\sum_{i \in U} \frac{D_{i,B}}{\pi_{i,B}} I_{(z_i \leq z)} I_{(x_i = x)}}{\sum_{i \in U} \frac{D_{i,B}}{\pi_{i,B}} I_{(x_i = x)}}. \quad (7)$$

Let $\hat{\Delta}_H^x$ be the uncertainty measure (4) with p.d.f.s $F_N(y|x)$ and $G_N(z|x)$ replaced by estimates (6) and (7), respectively. The following propositions hold.

Proposition 1. *Let $x \in \{x^1, \dots, x^K\}$, and suppose that*

$$\frac{\frac{n_A p(x)}{f_A(\zeta_A - 1)} - \frac{n_B p(x)}{f_B(\zeta_B - 1)}}{\frac{n_A p(x)}{f_A(\zeta_A - 1)} + \frac{n_B p(x)}{f_B(\zeta_B - 1)}} \rightarrow \alpha \text{ as } N \rightarrow \infty \quad (8)$$

Then, for almost all (x_i, y_i, z_i) s values, conditionally on X_N, Y_N, Z_N , as N increases:

$$\sqrt{\frac{\hat{n}_{H,A}(x)\hat{n}_{H,B}(x)}{\hat{n}_{H,A}(x) + \hat{n}_{H,B}(x)}}} \left(\hat{\Delta}_H^x - \Delta^x(F_N, G_N) \right) \xrightarrow{d} N(0, V(F, G; x)) \quad (9)$$

$N(0, V(F, G; x))$ denoting a normal r.v. with mean 0 and variance $V(F, G; x)$.

where

$$\lim_{N \rightarrow \infty} \frac{n_A}{N} = f_A, \quad \lim_{N \rightarrow \infty} \frac{n_B}{N} = f_B, \quad 0 < f_A < 1, \quad 0 < f_B < 1.$$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{\pi_{i,A}} = \zeta_A < \infty, \quad ; \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{\pi_{i,B}} = \zeta_B < \infty.$$

and

$$\hat{n}_{H,A}(x) = \frac{N \hat{p}_{H,A}(x)}{\frac{1}{N} \sum_{i=1}^N \pi_{i,A}^{-1} - 1}, \quad \hat{n}_{H,B}(x) = \frac{N \hat{p}_{H,B}(x)}{\frac{1}{N} \sum_{i=1}^N \pi_{i,B}^{-1} - 1}. \quad (10)$$

with $\widehat{p}_{H,A}(x)$ and $\widehat{p}_{H,B}(x)$ denoting the Hájek estimators of $p_N(x)$ obtained from the two (independent) samples A, B , respectively.

We now turn to the problem of estimating the unconditional measure of uncertainty (5). From the structure of (5), it is clear that the main problem we have to face is the estimation of the probabilities $p_N(x)$. The idea is use a estimator of $p_N(x)$ of the form

$$\widehat{p}_\tau(x) = \tau \widehat{p}_{H,A}(x) + (1 - \tau) \widehat{p}_{H,B}(x) \quad (11)$$

with $0 \leq \tau \leq 1$. As far as the value of τ is concerned, a fairly natural choice consists in taking the asymptotically optimal value of τ , *i.e.* the value of τ that minimizes the asymptotic variance of (11).

Proposition 2. *For almost all (x_i, y_i, z_i) s values, conditionally on X_N, Y_N, Z_N , as N increases the sequence of r.v.s $\sqrt{N}(\widehat{p}_\tau(x) - p_N(x))$ satisfies the relationship:*

$$\sqrt{N}(\widehat{p}_\tau(x) - p_N(x)) \xrightarrow{d} N(0, p(x)(1 - p(x))(\tau^2(\zeta_A - 1) + (1 - \tau)^2(\zeta_B - 1))). \quad (12)$$

As a consequence, the asymptotic variance of $\widehat{p}_\tau(x)$ is (proportional) to $p(x)(1 - p(x))(\tau^2(\zeta_A - 1) + (1 - \tau)^2(\zeta_B - 1))$, ad the asymptotically optimal value of τ is:

$$\tau^* = \frac{\zeta_A - 1}{\zeta_A + \zeta_B - 2}. \quad (13)$$

Hence, as an estimator of $p_N(x)$ we will take

$$\widehat{p}_{H,AB}(x) = \tau_N^* \widehat{p}_{H,A}(x) + (1 - \tau_N^*) \widehat{p}_{H,B}(x). \quad (14)$$

As a consequence of Proposition 2, it is immediate to see that the estimator (14) is asymptotically normally distributed with asymptotic variance $(\zeta_A - 1)(\zeta_B - 1)p(x)(1 - p(x))/(\zeta_A + \zeta_B - 2)$. This result can be extended to the (column) vector of estimates

$$\widehat{p}_{H,AB}(x) = [\widehat{p}_{H,AB}(x^1) \cdots \widehat{p}_{H,AB}(x^K)]' \quad (15)$$

where $x = [x^1 \cdots x^K]'$. Denote by $p_N(x)$ the (column) vector of elements $p_N(x^1), \dots, p_N(x^K)$, and let Σ be the $K \times K$ symmetric matrix of elements

$$\sigma_{hk} = \begin{cases} p(x^h)(1 - p(x^h)) & \text{if } k = h \\ -p(x^h)p(x^k) & \text{if } k \neq h \end{cases}; \quad h, k = 1, \dots, K. \quad (16)$$

We consider the following estimator of the unconditional measure of uncertainty (5)

$$\widehat{\Delta}_H = \sum_{k=1}^K \widehat{\Delta}_H^{x^k} \widehat{p}_{H,AB}(x^k) \quad (17)$$

its asymptotic normality is proved in proposition 3.

Proposition 3. *Let*

$$\hat{n}_A = \frac{n_A}{\frac{n_A}{N} \left(N^{-1} \sum_{i=1}^N \pi_{i,A}^{-1} - 1 \right)} = \frac{N}{N^{-1} \sum_{i=1}^N \pi_{i,A}^{-1} - 1}, \quad (18)$$

$$\hat{n}_B = \frac{n_B}{\frac{n_B}{N} \left(N^{-1} \sum_{i=1}^N \pi_{i,B}^{-1} - 1 \right)} = \frac{N}{N^{-1} \sum_{i=1}^N \pi_{i,B}^{-1} - 1}. \quad (19)$$

For almost all (x_i, y_i, z_i) s values, conditionally on X_N, Y_N, Z_N , as N increases the following result holds:

$$\sqrt{\frac{\hat{n}_A \hat{n}_B}{\hat{n}_A + \hat{n}_B}} \left(\hat{\Delta}_H - \Delta(F_N, G_N) \right) \xrightarrow{d} N(V(F, G)) \text{ as } N \rightarrow \infty \quad (20)$$

with

$$V(F, G) = \sum_{k=1}^K p(x) V(F, G; x) + \frac{(\zeta_A - 1)(\zeta_B - 1)}{(\zeta_A + \zeta_B - 2)^2} \Delta(F, G)' \Sigma \Delta(F, G). \quad (21)$$

References

1. Conti, P.L.: On the estimation of the distribution function of a finite populatrion under high entropy sampling designs, with applications. Submitted to Sankhya B.
2. Conti, P.L., Marella, D., Scanu, M.: Uncertainty Analysis in Statistical Matching. J. Off. Stat. **28**, 1–21 (2012)
3. Conti, P.L., Marella, D., Scanu, M.: Uncertainty Analysis for Statistical Matching of ordered categorical variables. Comput. Statist. Data Anal. **68**, 311–325 (2013)
4. Renssen, R.H.: Use of Statistical Matching Techniques in Calibration Estimation. Surv. Meth. **24**, 171–183 (1998)
5. Rubin, D.B.: Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. J. Bus. Econom. Statist. **4**, 87–94 (1986)