

Use of Auxiliary Information in Statistical Matching

L'Utilizzo dell'Informazione Ausiliaria nell'Integrazione Statistica

Rosalba Filippello, Ugo Guarnera

ISTAT, via Cesare Balbo 16, Roma - e-mail: filippel@istat.it

Giovanna Jona Lasinio

Università di Roma "La Sapienza"

Dipartimento di Statistica, Probabilità e Statistica Applicata

e-mail: giovanna.jonalasinio@uniroma1.it

Riassunto: Le tecniche di Integrazione Statistica, molto diffuse nel contesto delle Statistica Ufficiale, hanno l'obiettivo di comporre in un quadro coerente e complessivo informazioni provenienti da fonti di dati distinte. In questo lavoro è stato sperimentato, mediante uno studio di simulazione, un metodo di Integrazione Statistica basato sull'utilizzo di Informazione Ausiliaria valutandone le prestazioni attraverso un confronto con metodi classici basati sull'assunzione di Indipendenza Condizionata.

Keywords: Statistical Matching, Conditional Independence, Auxiliary Information

1. Introduction

Statistical inference from distinct data-sources is a relevant issue in the Official Statistics. The interest in this area arises from the need to investigate statistical properties of widely diffused data fusion procedures. Extracting information from different sources is necessary when one is interested in analyzing relationships between variables that are not jointly surveyed but resources for a new *ad hoc* survey are not available. In the framework of data integration two approaches can be distinguished: the *micro* approach and the *macro* approach (D'Orazio et al., 2002). The former aims at producing, through matching of units belonging to distinct sources, a unique data-file where for each unit all the variables are observed. On the contrary, the objective of the latter, is only the estimation of parameters or aggregates of the target population; in this case the production of a single data-set is not required. Inference from two distinct micro-files can be accomplished in two different situations. In the first, the two files consist of observations referring to the same units, although measurement errors or missing values might make difficult to match the corresponding units. The methodology dealing with such a situation is generally referred to as *record linkage* (Scanu, 2003). In the second case the two files are random samples from the same population, but only accidentally the same unit is sampled in both of them. In this situation, the methodology consisting in combining information from the two data sets, is known as *statistical matching* (Kadane, 2001). A typical situation in the framework of statistical matching can be described as follows. There are two sample files, file **R** and file **D**, taken from two different surveys. File **R** contains information on variables (X, Y) , whereas file **D** contains information on variables (X, Z) . The objective of statistical matching is to perform inferences on the joint distribution of (X, Y, Z) or (Y, Z) . These distributions cannot be

estimated on the base of the available information from file **R** and **D** alone unless it is assumed that variables **Y** and **Z** are independent conditional on **X**, i.e. $P(\mathbf{Y}, \mathbf{Z} | \mathbf{X}) = P(\mathbf{Y} | \mathbf{X}) \cdot P(\mathbf{Z} | \mathbf{X})$. In the absence of additional information on the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ or $(\mathbf{Y}, \mathbf{Z})$, most of statistical matching methods are based (more or less explicitly) on the *conditional independence assumption (CIA)*. In practice, under this assumption, if a single “complete” micro file is to be produced, some similarity criterion (generally based on a distance function) is defined in terms of the common variables **X** and “similar” units from files **R** and **D** are matched to create complete records. At macro level, the CIA implies that $P(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ can be recovered by estimating $P(\mathbf{Y} | \mathbf{X})$ from file **R**, $P(\mathbf{Z} | \mathbf{X})$ from file **D** and the marginal distribution $P(\mathbf{X})$ from both **R** and **D**. It is clear that departures from conditional independence will determine serious bias in the estimates of the distributions $P(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ and $P(\mathbf{Y}, \mathbf{Z})$. The strong assumption of CI can be avoided if *auxiliary information (AI)* on the $(\mathbf{Y}, \mathbf{Z})$ or $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ relationships is available in the form of a third micro file **C** containing the full set of variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ or the reduced set $(\mathbf{Y}, \mathbf{Z})$ (Paass, 1986; Singh et al., 1993). The role of AI is mainly focused in reducing the bias in the estimates of the joint distributions. AI should not affect the estimates that can be based on the files **R** and **D** only. In particular, the only estimates whose uncertainty depends on the sampling variability of the AI are those referring to parameters associated with $(\mathbf{Y}, \mathbf{Z})$ or $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ relationships.

2. Simulation Study

In this paper we describe a method of statistical matching based on the use of auxiliary information through a simulation study, and compare it with three methods based on the CIA. Data for files **R** and **D** are taken from the April 2001 Labour Force Survey (ISTAT), while we use historical data (1999, 2000) from the same survey as AI. All the considered data-sets contain about 5000 observations. The observations refer to employees of Toscana region. In our study we consider data as realisations of *i.i.d.* variables from an underlying superpopulation model. In such a way we do not take into account the sampling design. The variables have been chosen as follows. The common variable **X** is the “concatenation” of variables *sex* and *age class* (i.e. **X** categories are obtained by crossing *sex* and *age class* categories), **Y** is the *sector of economic activity* and **Z** is the *professional position*. For all the methods to be evaluated 100 simulations have been run. In each simulation the statistical matching situation has been mimed by randomly splitting the original data-set into two subsets of data **R**, **D** of about the same size and “obscuring” **Z** and **Y** in file **R** and **D** respectively. Our (*macro*) objective was to evaluate the different methods in terms of their ability to reproduce the counts n_{ijk} of the original contingency table $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ where indices $i=1, \dots, I$; $j=1, \dots, J$; $k=1, \dots, K$ run over the categories of **X**, **Y**, **Z** respectively. Three CIA-based methods have been applied. The first (DON1) consisted in imputing synthetic values for **Z** in file **R** and for **Y** in file **D** by a random donor technique within matching classes defined through the common variable **X**. A second version of the donor technique (DON2) has been also applied based on a distance function in the **X**-space. To this aim a larger number of **X** variables have been used and each matching variable has been weighted with an index measuring its association with the variable (**Y** or **Z**) to be imputed. The third method (LOGLIN) is parametric. The estimates of the “cell probabilities” $\pi_{ijk} \equiv P(X=X^i, Y=Y^j, Z=Z^k)$ are determined via the expression $\pi_{ijk} \equiv \pi_{j|i} \cdot \pi_{i \cdot \cdot} \cdot \pi_{k|i}$ where $\pi_{j|i} \equiv P(Y=Y^j | X=X^i)$ are

estimated from file **R**, $\pi_{k|i} \equiv P(Z=Z^k | X=X^i)$ from file **D**, and marginal probabilities $\pi_{i.} \equiv P(X=X^i)$ from the whole data-set obtained by concatenating file **R** and **D**. It is worthwhile noting that, considering files **R** and **D** as independent random samples from the same population, allows interpreting the above procedure as the maximum likelihood estimate (MLE) of the parameters of a log-linear model corresponding to conditional independence of Y and Z given X . More precisely, the estimates are determined as MLE's based on the "incomplete data" likelihood function resulting from jointly considering the observations of files **R** and **D** under a loglinear model with the interaction terms λ^{YZ} and λ^{XYZ} set to zero.

In the last group of experiments auxiliary information has been taken from historical data. In order to study the impact of the sample size of the auxiliary information on the method performance, we carried out ten groups of simulations varying the number of units in the historical file. In each group of simulations we extracted from available information simple random samples of fixed size, with sizes ranging from about 10% up to 100% of the whole historical data-set. With the aid of AI we tried to estimate the parameters of a saturated log-linear (multinomial) model, i.e. the cell probabilities $\pi_{ijk} \equiv P(X=X^i, Y=Y^j, Z=Z^k)$. Due to the lack of joint information on (X, Y, Z) the multinomial model cannot be estimated on the basis of file **R** and **D** only. This implies that the ML equations for π_{ijk} do not have a unique solution. Nevertheless, they determine univocally the ML estimates of the marginal probabilities $\hat{\pi}_{ij.}, \hat{\pi}_{i.k}$ that are based on the sample frequencies $n_{ij.}$ and $n_{i.k}$ (Agresti, 1990). In our approach, we considered these estimates as constraints for the parameters $\pi_{ijk.}$. The parameters π_{ijk} were estimated by Iterative Proportional Scaling (IPS) of the marginal probabilities $\hat{\pi}_{ij.}, \hat{\pi}_{i.k}$ on a set of frequencies obtained in two different ways:

AI1) using sample frequencies n_{ijk} in the historical file.

AI2) uniformly spreading historical (Y, Z) marginals $n_{.jk}$ over the X categories.

With the second procedure we pretend only auxiliary information on the (Y, Z) distribution is available.

3. Results and conclusions

In this paper different methods have been evaluated by comparing the cell frequencies of the original data-set with the corresponding frequencies resulting from the statistical matching process. In particular we use a χ^2 test for testing whether two set of sample frequencies can be considered as if they come from the same distribution (Mood et al. 1997). The p -values from this test have been considered as performance indices and evaluations have been performed by analyzing the corresponding distributions over the 100 simulations for each method. The analysis of the results shows that:

- 1) Marginal univariate distributions and bivariate (X, Y) , (X, Z) distributions are well preserved by all the methods.
- 2) All the CIA-based methods have poor performances in recovering (Y, Z) and (X, Y, Z) distributions (all p -values below 0.05).
- 3) DON1 and DON2 yield more variable results with respect to LOGLIN due to the extra variability introduced by the donor technique.
- 4) Methods based on auxiliary information show increasingly better performances as the sample size of the historical file increases (all p -value greater than 0.05 starting from a sample size of about 2000 observations).

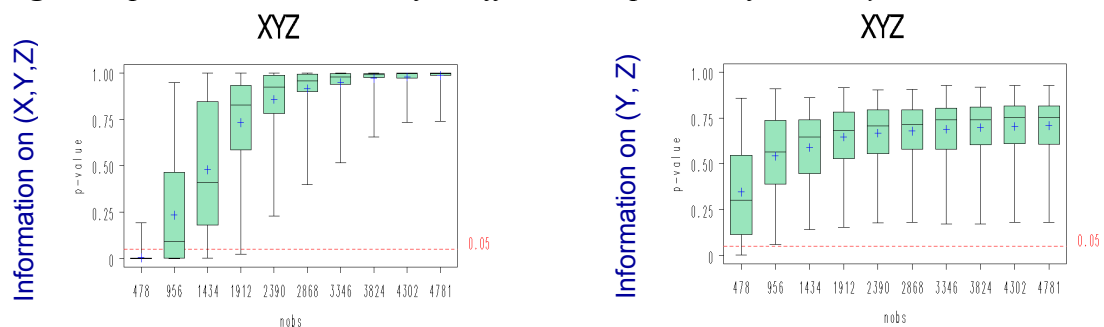
5) For small sample sizes (less than 1000 observations) of the auxiliary information, the (Y, Z) distribution is preserved better than the (X, Y, Z) distribution by the statistical matching process.

6) For small sample sizes of the auxiliary information, using information on (Y, Z) distribution yields better results than using information on the full set of variables (X, Y, Z) (Figure 1).

From the above results it is clear that in our case the CIA is inadequate if one is interested in analysing relationships between not jointly observed variables (Y, Z) . Furthermore, it can be argued that, if CI has to be assumed because of the lack of auxiliary information, a more simple parametric method is to be preferred to a computationally demanding method such as the donor-based technique.

As far as the use of auxiliary information is concerned, an interesting finding seems to be contained in result 6): the somewhat surprising result that, with small sample size, using less information gives better results than using all the available information could be due to the instability in the estimation of the (X, Y, Z) factor effects on account of the insufficient size of auxiliary data. Thus, if this effect is small, setting it to zero is perhaps better than trying to estimate it with modest additional information.

Figure 1: *p-values distributions for different sample size of auxiliary dataset.*



Acknowledgments: the authors would like to thank P.L. Conti, M. Di Zio and M. Scanu for helpful comments and suggestions.

References

- Agresti, A. (1990) *Categorical Data Analysis*. Wiley & Sons, New York.
- D'Orazio, M., Di Zio, M., Scanu, M. (2002), Statistical Matching and Official Statistics, *Quaderni di Ricerca ISTAT*, 1, 5-24.
- Kadane, J.B. (2001) Some Statistical Problems in Merging Data Files, *Journal of Official Statistics*, 17, 423-433.
- Mood, A.M., Graybill, F.A., Boes, D.C. (1997) *Introduzione alla Statistica*, McGraw – Hill Italia Srl.
- Paass, G., (1986) Statistical Match: Evaluation of Existing Procedures and Improvements by using Additional Information, *Microanalytic Simulation Models to Support Social and Financial Policy*, (Eds. G.H. Orcutt, J. Merz and H. Quinke), Elsevier Science Publishers B.V. (North-Holland - Amsterdam).
- Scanu M. (2003) Metodi Statistici per il Record Linkage *Metodi e Norme ISTAT*, 16.
- Singh, A.C., Mantel, H., Kinack, M., Rowe, G. (1993) Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption, *Survey Methodology*, 19, 59-79.