Old and new approaches in statistical matching when samples are drawn with complex survey designs

Marcello D'Orazio, Marco Di Zio, Mauro Scanu

Abstract Statistical matching tackles the problem of drawing information on a pair of random variables (Y,Z) which have not been observed jointly in one sample survey. In fact, Z and Y are available in two distinct and independent surveys whose sets of observed units are non overlapping. The two surveys observe also some common variables X. This problem has traditionally been analyzed when the two sample surveys consists of independent and identically distributed observations from the same model. On the contrary most surveys, especially those managed in National Statistical Institutes, consists of samples drawn from a finite population according to complex survey designs. This paper compares some of the procedures described in the literature and analyzes their effects on the analysis of uncertainty, *i.e*. of the lack of joint information on the random variables of interest.

Key words: File concatenation, calibration, empirical likelihood, identifiability.

1 What is statistical matching

Statistical matching techniques [3] combine information available in distinct data sources referred to the same target population. The two datasets, A and B, are assumed to contain data collected in two independent sample surveys and such that

- the two samples contain distinct units (the sets of units observed in A and B do not overlap);

Marcello D'Orazio
Istat, via Cesare Balbo 16 - 00184 Roma, e-mail: madorazi@istat.it

Marco Di Zio
Istat, via Cesare Balbo 16 - 00184 Roma, e-mail: dizio@istat.it

Mauro Scanu
Istat, via Cesare Balbo 16 - 00184 Roma, e-mail: scanu@istat.it

- the two samples contain information on some variables X (common variables), while other variables are observed distinctly in one of the two samples, say, Y in A and Z in B.

In statistical matching, the key problem is the relationship among the variables Y and Z, given that they are never jointly observed in the data sets at hand. Analysts have always questioned if it was possible to either create synthetic records containing information on (X,Y,Z), as in [11], or through inference on model parameters describing the joint behaviour of the variables, as correlation coefficients [6, 14, 9, 10, 12], conditional probabilities [13, 2], and so on.

The model is not identifiable given the available data, unless some untestable models are assumed, as conditional independence between Y and Z given X. Most of the literature on this topic is based on the conditional independence assumption. A more truthful study would limit inference on "how unidentifiable" is the model: this problem leads to the analysis of "uncertainty", as suggested by [12] and [3]. In this context, an inferential problem on (X,Y,Z) does not end up with a punctual estimate for the target quantities (function of the variable of interests), but, broadly speaking, consists of an interval that encloses all the possible values coherent with the observed information (*e.g.* an interval for a correlation coefficient of the variables "household expenditures" observed in a survey and "household consumption" observed in a second survey). We remark that this interval is different from the inference based on confidence intervals, in the latter the uncertainty taken into account is due to sampling variability, in this case the uncertainty is due to the lack of information that implies model unindentifiability [8]. The computation of an interval instead of a single punctual value is inherently related to the absence of joint information on the variables observed distinctly in the two sample surveys. Papers in this context provide algorithms to build those intervals both in case of Gaussian distributions [9, 12, 7] and of multinomial distributions [2], however they refer to independent and identically distributed (iid) samples. In National Statistical Institutes (NSIs), survey data to be integrated generally refer to samples selected from the same finite population through a complex sampling design. There are not many studies concerning this issue [14, 13], and they are not directly related to the analysis of "uncertainty".

This paper is devoted to the study of "uncertainty" in statistical matching in the framework of complex sampling designs. Without losing in generality, this paper deals only with categorical variables.

2 Statistical matching for samples drawn according to complex survey designs

When samples are drawn according to complex survey designs [14] and [13] are the two main references for statistical matching. A third approach given in [16], although not explicitly designed for statistical matching, can be used also in this

context. In order to assess the uncertainty of statistical matching, it is enough to illustrate these approaches under the assumption of conditional independence of Y and Z given X.

Before showing the various approaches it is worth introducing some notations. Let U be the finite population of size N. Let us consider the random sample A selected from U with the sampling design $p(A)$ consisting of n_A sample units and the random sample B selected from U with the sampling design $p(B)$ consisting of n_B sample units. Let $d_{A,a} = 1/\pi_{A,a}$ be the direct weight associated to each sample unit in A, and $d_{B,b} = 1/\pi_{B,b}$ be the corresponding direct weight for the units in B. The variables (X,Y) are observed in A while (X,Z) are observed in B.

2.1 File concatenation

The original proposal of [14] consisted in modifying the sample weights of the two surveys A and B in order to get a unique sample given by the union of A and B ($A \cup B$) with new survey weights representative of the population of interest. The basic idea is that new sampling weights can be derived from the concatenated files by using the simplifying assumption that the probability of including a unit in both the samples is negligible. This is generally true when the two sample surveys are independent and when the sampled fraction of the population in the two cases is negligible, as in many social surveys. Under this assumption, the inclusion probability of a record $c \in A \cup B$ is:

$$\pi_c = \pi_{A,c} + \pi_{B,c} \ . \tag{1}$$

These new weights can be used in order o estimate the parameters of interest in the concatenated file A *bigcup*B.

Note however that for each unit in $A \cup B$ the inclusion probability of the records in A under the survey design in B, as well as the inclusion probability of the records in B under the survey design in A, must be computed. It is worth noting that design variables of a survey are not necessarily available in other surveys and for this reason the approach proposed by Rubin has been seldom applied.

If the probability of inclusion of a unit in both samples is not negligible (*e.g.* this happens when survey designs admit take all strata) Equation 1 becomes :

$$\pi_c = \pi_{A,c} + \pi_{B,c} - \pi_{A \cap B,c} \ . \tag{2}$$

2.2 Approaches that use the surveys separately

This approach was described at first in [13]. This approach has a much wider applicability, because it is not necessary to know the design variables of both surveys. As a matter of fact, the two different samples are preserved and their statistical content is harmonised by calibrating the two sets of survey weights. The result is the creation

of two samples that are able to estimate consistently totals of the common information. Calibration follows one (or both) of the following two criteria, depending on the characteristics of the common variables

1. if a common variable is known on the whole population, calibration should reproduce the known population frequencies;
2. if a common variable is not known on the whole population, an estimate is given by a linear combination of the estimates computed in A and B respectively, and calibration should reproduce this final estimate.

When dealing with the multivariate common information (X) it has to be decided whether to preserve just the marginal distributions of each component of X or their joint distribution. Constraining on the joint distribution poses some computational problems.

Under the conditional independence assumption, estimation is performed by estimating the common variables X either in A or B, the conditional distribution of Y given X in A and the conditional distribution of Z given X in B.

A similar approach is introduced in [16]. This approach is based on the pseudo empirical likelihood but the derivation of the new sets of survey weights follows a different procedure that forbids negative final weights. This approach is known as Wu separate approach. Wu [16] suggests also an alternative approach, known as Wu combined approach. This approach avoids that the distribution of X is estimated in advance by simply introducing a constraint that the new sets of weights reproduce the same X distribution both in A and B.

2.3 Empirical likelihood

The empirical likelihood in the context of statistical matching is described in detail in [4]. Consider the log-likelihood function

$$l(p|A\bigcup B) = \sum_{s=1}^{N} \log(p_X(x_s)) + \sum_{s=1}^{N} \log(p_{Y|X}(y_s|x_s)) + \sum_{s=1}^{N} \log(p_{Z|X}(z_s|x_s)). \quad (3)$$

If the entire finite population was available, our objective would be to maximize (3) with respect to p. In practice we have a sample and thus, as suggested by [1], since the log-likelihood (3) can be understood as a sum of finite population totals, our goal becomes to maximize the estimate $\hat{l}(p|A\bigcup B)$ (the pseudo empirical log-likelihood) with respect to the vector of parameters p: these are the pseudo empirical likelihood (MPEL) estimates.

The estimates that can be obtained through the approaches described in Section 2 can be justified also as MPELs. The differences concern how each component of (3) is estimated.

File concatenation. - As suggested in Section 2.1, concatenate the samples A and B, compute a unique set of weights $d_c = 1/\pi_c$, $c \in A\bigcup B$, as in Equation (2),

and use those weights to make inference on the concatenated file by means of the following estimate of (3):

$$\widehat{l}(p|A\bigcup B) = \sum_{c\in A\bigcup B} d_c log(p_X(x_s)) + \sum_{a\in A} d_a \log(p_{Y|X}(y_s|x_s)) + \sum_{b\in B} d_b \log(p_{Z|X}(z_s|x_s))$$

Following Rubin's suggestion, the survey weights d_c should be modified so that their sum is equal to N. Note that estimators of the conditional distributions of Y given X and Z given X should use modified weights d_a^*, $a \in A$, and d_b^*, $b \in B$, which are computed in order to take missing data into account.

Approaches that use the surveys separately. In this case, the estimate of equation (3) is:

$$\begin{aligned}\widehat{l}(p|A\bigcup B) = &\sum_{a=1}^{n_A} d_{A,a} \log(p_X(x_a)) + \sum_{b=1}^{n_B} d_{B,b} \log(p_X(x_b)) \\ &+ \sum_{a=1}^{n_A} d_{A,a} \log(p_{Y|X}(y_a|x_a)) + \sum_{b=1}^{n_B} d_{B,b} \log(p_{Z|X}(z_b|x_b)). \end{aligned} \quad (4)$$

The goal is to maximize this pseudo empirical log-likelihood. Renssen, as suggested in 2.2, obtains a unique estimate of the common variable X by combining the estimates of the frequency distribution of X from A and B, and then calibrates the original sample weights $d_{A,a}$ and $d_{B,b}$ to the obtained estimate $\widehat{p}_X(x)$. The new weights are used to estimate the distributions of $Y|X$ in A and $Z|X$ in B. These estimates can be considered as MPEL estimates on A and B respectively.

Wu separate approach differs from the Renssen's one for the derivation of the final sets of weights $d_{A,a}$ and $d_{B,b}$.

Wu combined approach uses the constraint that the two estimates of the X distribution from A and B respectively are the same. Details can be found in [16].

3 Uncertainty

As remarked in D'Orazio et al. (2006a), the maximum likelihood estimates may be used to determine the impact of the absence of joint information on Y and Z on the estimates of the joint (Y,Z) parameters. This can be described in terms of an interval of estimates that are equally plausible for the available data sets. This interval can be estimated by the likelihood ridge. This consideration immediately applies when samples are drawn according to complex survey designs when a MPEL approach is used.

More precisely, let the distributions satisfying the system of equations:

$$\Theta = \begin{cases} \sum_k p_{XYZ}(i,j,k) = p^*_{XY}(i,j) & i = 1,\ldots,I,\ j = 1,\ldots,J \\ \sum_j p_{XYZ}(i,j,k) = p^*_{XZ}(i,k) & i = 1,\ldots,I,\ k = 1,\ldots,K \\ p_{XYZ}(i,j,k) \geq 0;\ \sum_{ijk} p_{XYZ}(i,j,k) = 1 & \end{cases} \quad (5)$$

be the uncertainty on $p_{XYZ}(i,j,k)$ when the distributions $p_{XY}(i,j) = p^*_{XY}(i,j)$ and $p_{XZ}(i,k) = p^*_{XZ}(i,k)$ are known.

An estimate of Θ can be obtained by all the MPEL estimates obtainable from A and B according to any of the schema shown in the previous section

$$\widehat{\Theta} = \begin{cases} \sum_k p_{XYZ}(i,j,k) = \widehat{p}_{XY}(i,j) & i = 1,\ldots,I,\ j = 1,\ldots,J \\ \sum_j p_{XYZ}(i,j,k) = \widehat{p}_{XZ}(i,k) & i = 1,\ldots,I,\ k = 1,\ldots,K \\ p_{XYZ}(i,j,k) \geq 0;\ \sum_{ijk} p_{XYZ}(i,j,k) = 1 & \end{cases} \quad (6)$$

For each single parameter, the pseudo empirical likelihood ridge can be described by the use of the Fréchet class, where the components are the corresponding MPEL estimates of $p_X(i)$, $p_{Y|X}(j|i)$ and $p_{Z|X}(k|i)$.

For the bivariate distribution of (Y,Z) it is:

$$\sum_i \widehat{p}_X(i) \max\{0; \widehat{p}_{Y|X}(j|i) + \widehat{p}_{Z|X}(k|i) - 1\} \leq p_{YZ}(jk) \leq \sum_i \widehat{p}_X(i) \max\{\widehat{p}_{Y|X}(j|i); \widehat{p}_{Z|X}(k|i)\}.$$

4 Simulation results

The performances of the estimators of $p_X(x)$, $p_{Y|X}(y|x)$ and $p_{Z|X}(z|x)$ under the different schema are evaluated using a simulation study. The finite population used in this study consists of an artificial population U of $N = 5000$ individuals with age greater than 15 years and being occupied in a dependent position. The following variables are considered: Geographical Area (grouped respectively in 3 or 5 categories); Gender, X_1, (1='M', 2='F'); Classes of Age, X_2, (3 classes: '16–22', '23–44', '$\geq$ 45'); Education Level, Y, (4 categories: 1='No title or elementary school', 2='compulsory school', 3='Secondary school', 4='university degree or higher'); and Professional Status, Z, (3 categories: 1='worker', 2='employee', 3='manager'). In order to reproduce the statistical matching framework and then use the approaches presented in the previous sections to estimate $p_X(x)$, $p_{Y|X}(y|x)$ and $p_{Z|X}(z|x)$, in each simulation run two random samples A and B are selected from U. The samples are selected using a stratified random sampling with proportional allocation; Geographical Area is used as the stratification variable. Three strata ('North', 'Center' and 'South and Islands') are considered when selecting the sample A, while the strata are five ('North West', 'North East', 'Center', 'South' and 'Islands') when selecting B. As far as the sampling fractions are concerned, three combinations are considered:

(a) $f_A = 0.10$ ($n_A = 500$) and $f_B = 0.06$ ($n_B = 300$);
(b) $f_A = f_B = 0.10$ ($n_A = n_B = 500$);
(c) $f_A = 0.06$ ($n_A = 300$) and $f_B = 0.10$ ($n_B = 500$).

In sample A the variable Z is removed, as Y in file B. The whole process is repeated $T = 10000$ times for each combination of the sampling fractions.

Tables 1, 2 and 3 report the averages, over the whole set of simulations, of the absolute distance (total variation distance) among the estimated and the true population relative frequencies ($p_c = N_c/N$) computed using

$$\bar{d}(\hat{p}, p) = \frac{1}{10000} \sum_{t=1}^{10000} \left[\frac{1}{2} \sum_{c=1}^{C} |\hat{p}_{t,c} - p_c| \right] \tag{7}$$

where $\hat{p}_{t,c}$ is the estimate of $p_c = N_c/N$ computed on the t-th run, $t = 1, \ldots, 10000$.

Table 1 Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (a).

Approach	$p_X(x)$	$p_{XY}(x,y)$	$p_{Y\|X}(y\|x)$	$p_{XZ}(x,z)$	$p_{Z\|X}(z\|x)$
File concatenation	2.617663	6.130204	5.494019	6.667832	5.494019
Renssen	2.617715	6.130005	5.493756	6.667069	5.493756
Wu separate	2.617715	6.130005	5.493756	6.667069	5.493756
Wu combined	2.795445	6.219056	5.493756	6.852094	5.493756

Table 2 Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (b).

Approach	$p_X(x)$	$p_{XY}(x,y)$	$p_{Y\|X}(y\|x)$	$p_{XZ}(x,z)$	$p_{Z\|X}(z\|x)$
File concatenation	2.621003	7.746804	7.264135	5.382213	7.264135
Renssen	2.620999	7.747073	7.264329	5.382442	7.264329
Wu separate	2.620999	7.747073	7.264329	5.382442	7.264329
Wu combined	3.037953	7.894812	7.264329	5.487439	7.264329

Table 3 Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (c).

Approach	$p_X(x)$	$p_{XY}(x,y)$	$p_{Y\|X}(y\|x)$	$p_{XZ}(x,z)$	$p_{Z\|X}(z\|x)$
File concatenation	2.621003	7.746804	7.264135	5.382213	7.264135
Renssen	2.620999	7.747073	7.264329	5.382442	7.264329
Wu separate	2.620999	7.747073	7.264329	5.382442	7.264329
Wu combined	3.037953	7.894812	7.264329	5.487439	7.264329

The proposed estimation schema provide quite close results. In general, Renssen approach and Wu separate provide the same results. As expected, when estimating $p_X(x)$ file concatenation is slightly better than the other ones, but the differences with Renssen approach and Wu separate are really negligible. In this case Wu combined approach gives always the worst results. As far as the joint or the conditional distributions are concerned, there are not manifest patterns in the results; results under file concatenation are slightly better in situations (b) and (c); on the contrary Renssen approach and Wu separate perform slightly better in case (a). In all the

cases, the distance between true and estimated proportions tends to increase when estimation is carried out in the smaller sample.

Tables 4, 5 and 6 show summary results related to the estimation of uncertainty bounds for $p_{XY}(x,y)$ starting from the estimates obtained for $p_X(x)$, $p_{Y|X}(y|x)$ and $p_{Z|X}(z|x)$ at the end of each simulation under the different estimation schema. In particular, the values in the tables refer to the average width of the intervals built using uncertainty bounds in the simulations:

$$\overline{R}_c^{(low)} = \frac{1}{10000}\sum_{t=1}^{10000}\left(\hat{p}_{t,c}^{(up)} - p_{t,c}^{(low)}\right), \qquad c = 1,\ldots,C \tag{8}$$

Table 4 Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (a).

Edu	Prof	pop. prop.	File conc.	Renssen	Wu sep.	Wu comb
1	1	0.0484	0.050842	0.050809	0.050809	0.050812
2	1	0.2806	0.320464	0.320439	0.320439	0.320445
3	1	0.1536	0.383846	0.383815	0.383815	0.383827
4	1	0.0074	0.133076	0.133091	0.133091	0.133090
1	2	0.0024	0.050839	0.050806	0.050806	0.050809
2	2	0.0536	0.304700	0.304697	0.304697	0.304707
3	2	0.2810	0.372366	0.372359	0.372359	0.372375
4	2	0.0872	0.133096	0.133110	0.133110	0.133110
1	3	0	0.043179	0.043140	0.043140	0.043140
2	3	0.0060	0.085776	0.085704	0.085704	0.085707
3	3	0.0412	0.085780	0.085709	0.085709	0.085711
4	3	0.0386	0.073677	0.073653	0.073653	0.073652

Table 5 Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (b).

Edu	Prof	pop. prop.	File conc.	Renssen	Wu sep.	Wu comb
1	1	0.0484	0.050798	0.050833	0.050833	0.050840
2	1	0.2806	0.322623	0.322637	0.322637	0.322638
3	1	0.1536	0.387250	0.387273	0.387273	0.387272
1	1	0.0074	0.133083	0.133074	0.133074	0.133076
2	2	0.0024	0.050795	0.050829	0.050829	0.050837
3	2	0.0536	0.306699	0.306690	0.306690	0.306692
1	2	0.2810	0.376183	0.376180	0.376180	0.376186
2	2	0.0872	0.133087	0.133078	0.133078	0.133080
3	3	0	0.044193	0.044227	0.044227	0.044230
1	3	0.0060	0.085666	0.085707	0.085707	0.085702
2	3	0.0412	0.085666	0.085707	0.085707	0.085702
3	3	0.0386	0.074583	0.074595	0.074595	0.074591

The average width of the cell proportions computed by considering the estimates of the Fréchet bounds remains essentially the same under the various schema. Again, file concatenation tends to perform slightly better in cases (b) and (c). Slightly better results are obtained under Renssen approach or Wu separate in case (a).

Table 6 Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (c).

Edu	Prof	pop. prop.	File conc.	Renssen	Wu sep.	Wu comb
1	1	0.0484	0.050607	0.050637	0.050637	0.050626
2	1	0.2806	0.320618	0.320633	0.320633	0.320683
3	1	0.1536	0.383861	0.383870	0.383870	0.383917
1	1	0.0074	0.132868	0.132848	0.132848	0.132845
2	2	0.0024	0.050603	0.050633	0.050633	0.050622
3	2	0.0536	0.304739	0.304741	0.304741	0.304784
1	2	0.2810	0.372025	0.372028	0.372028	0.372069
2	2	0.0872	0.132882	0.132861	0.132861	0.132858
3	3	0	0.042914	0.042940	0.042940	0.042939
1	3	0.0060	0.085592	0.085615	0.085615	0.085611
2	3	0.0412	0.085598	0.085621	0.085621	0.085617
3	3	0.0386	0.073646	0.073650	0.073650	0.073648

To conclude, the results of this simulation experiment highlight that there is not an approach that outperforms the other ones. File concatenation seems to provide slightly better results than Renssen approach and Wu separate. The one that tends to perform slightly worst than the other ones is Wu combined.

5 Conclusions

This paper provides a unified view of some approaches to apply statistical matching when dealing with data sources arising from complex survey sampling from the same finite population. This is permitted by considering the pseudo empirical likelihood approach to finite population sampling.

The various schema proposed provide essentially the same results in terms of estimation of the target distributions and of the uncertainty bounds. Renssen and Wu separate approaches are really similar and even if they use a different method to compute the final weights, not surprisingly [16] they provide the same results. Both the approaches are based on a subjective decision concerning the estimation of the distribution $p_X(x)$ as a linear combination of the estimates obtained from A and B separately. Wu separate approach, in any case seems more flexible, because it does not allow for negative survey weights as it is possible when linear calibration is considered. The decision concerning the reference distribution for $p_X(x)$ can be avoided by considering Wu combined approach but, in this case, results obtained are slightly worst than the other ones.

In general the approaches suggested by Wu appear to be more flexible if compared to standard calibration approach. Anyway, their application requires much care because in presence of a stratified sampling design (not with proportional allocation) the theory is more complex and higher computational effort is required.

Rubin file concatenation provides quite good results, sometimes slightly better than the ones under the other approaches. It is the best as far as estimation of $p_X(x)$ is concerned, due to the fact that this estimate is carried out on $A \bigcup B$. Moreover,

in this step there is not the problem of harmonizing the estimates of $p_X(x)$ coming from A and B. Unfortunately, Rubin's approach is based on the computation of the concatenated weights, this may not be possible in some situations. Moreover, once computed the concatenated weights, a method to estimate joint or conditional distributions (i.e. $p_{XY}(x,y)$ or $p_{Y|X}(y|x)$) in presence of missing values should be chosen.

References

1. Chen, J., Sitter, R.R.: A pseudo empirical likelihood approach to the effctive use of auxiliary information in complex surveys. Stat. Sinica **9**, 385–406 (1999)
2. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical matching for categorical data: Displaying uncertainty and using logical constraints. JOS **22**, 137–157 (2006a)
3. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical matching: Theory and practice. Wiley, Chichester (2006b)
4. D'Orazio, M., Di Zio, M., Scanu, M.: Uncertainty intervals for nonidentifiable parameters in statistical matching. Proceedings 54th ISI session, 16-22 August 2009, Durban, South Africa
5. Fattorini, L.: Applying the Horvitz-Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. Biometrika **93**, 269–27 (2006)
6. Kadane, J.B.: Some statistical problems in merging data files. In: Compendium of tax research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179 (1978). Reprinted in 2001, JOS **17**, 423–433
7. Kiesl, H., Rässler, S.: The Validity of Data Fusion. CENEX-ISAD workshop, Wien 29-30 May 2008. Available at http://cenex-isad.istat.it
8. Manski, C.F.: Identification problems in the social sciences. Harvard University Press, Cambridge, Massachusetts (1995)
9. Moriarity, C., Scheuren, F.: Statistical matching: a paradigm for assessing the uncertainty in the procedure. JOS **17**, 407–422 (2001)
10. Moriarity, C., Scheuren, F.: A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation. J. Bus. Econ. Stat. **21**, 65–73 (2003)
11. Okner, B.A.: Constructing a new data base from existing microdata sets: the 1966 merge file. Ann. Econ. Soc. Meas. **1**, 325342 (1972)
12. Rässler, S.: Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York (2002)
13. Renssen, R.H.: Use of statistical matching techniques in calibration estimation. Survey Methodology **24**, 171–183 (1998)
14. Rubin, D.B.: Statistical matching using file concatenation with adjusted weights and multiple imputations. J. Bus. Econ. Stat. **4**, 87–94 (1986)
15. Torelli N., Ballin M., D'Orazio M., Di Zio M., Scanu M., Corsetti G.: Statistical matching of two surveys with a nonrandomly selected common subset. CENEX-ISAD workshop, Wien 29-30 May 2008. Available at: http://cenex-isad.istat.it
16. Wu, C.: Combining information from multiple surveys through the empirical likelihood method. Can. J. Stat. **32**, 1-12 (2004)