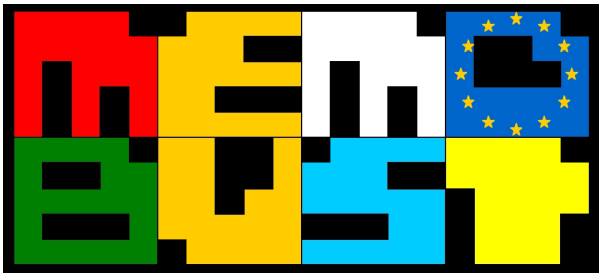

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Matching

Contents

General section

1. Summary

This section explores the problem of data integration in the following context: there are two non-overlapping surveys (in the sense that the two sets of units collected in the two surveys are distinct) that refer to the same target population, the variables of interest for the statistical analyses are available distinctly in the two surveys, due to the nature of the data sets it is not possible to create joint information on these variables by means of their common identifiers. This problem is usually referred to as statistical matching. As a matter of fact, this is a non-standard problem in statistics, for which naïve methods based on data imputation were defined at the beginning. Nowadays the complex nature of statistical matching is dealt differently, by the exploration of all the possible models that could give as a result the two sample surveys at hand, giving rise to "sets" of estimates instead of the more usual "point estimates". These sets of estimates should not be confused with confidence intervals: they just reflect the fact that joint information on the target variables is missing.

2. General description

Statistical matching (sometimes called data fusion, synthetical matching) aims at combining information available in distinct sample surveys referred to the same target population. Formally, let Y and Z be two random variables (r.v.). Statistical matching is defined as the estimation of the joint (Y,Z) distribution function (e.g., a contingency table or a regression coefficient) or of some of its parameters when:

- Y and Z are not jointly observed in a survey, but
- Y is observed in a sample A, of size n_A,
- Z is observed in a sample B, of size n_B,
- A and B are independent, and the set of observed units in the two samples do not overlap (it is not possible to use record linkage),
- A and B both observe a set of additional variables X.

A figure representing this situation is the following.

	Y	**X**	**Z**
Data source *A*			missing
	Y	**X**	**Z**
Data source *B*	missing		

A detailed list of statistical matching applications is in D'Orazio et al. (2006) and Ridder and Moffit (2007). Generally speaking, this problem has been considered as an imputation problem. One of the

files, e.g., A, was considered the recipient, the other the donor file, and the statistical matching procedure consists in imputing Z in A by means of the available common information X. Among the procedures applied in this context, it is possible to distinguish

1. Use of imputation techniques that reproduce the assumption of independence of Y and Z given X (conditional independence assumption, henceforth CIA). One of the first statistical matching attempts is in Okner (1972). In this case, statistical matching consisted of the application of imputation techniques of taxable income observed on 1966 Internal Revenue Service Tax File on the 1967 Survey of Economic Opportunity. Denoting the common variables in the two files as X, the variables observed only in the Survey of Economic Opportunity as Y and those only in the Tax File as Z, these imputation techniques were able to reproduce the model of conditional independence between Y and Z given X. Appropriateness of CIA is discussed in several papers. We quote, among the others, Sims (1972) and Rodgers (1984).
2. Use of external auxiliary information for avoiding the CIA. This second group of techniques uses external auxiliary information on the statistical relationships between Y and Z, e.g., an additional file C where (X, Y, Z) are jointly observed is available (as in Singh et al., 1993).

The imputation procedures used in the two previous contexts can be clustered in:

1. parametric: i.e., explicit use of a parametric model (e.g., a regression) between X, Y and Z
2. nonparametric: use of hot-deck methods
3. mixed: two step procedures that partially make use of parametric models and then apply hot-deck methods for imputation of "live" values

These approaches are actually theoretically justified when the joint probability distribution of the variables of interest in the population coincides with the probability distribution of the same variables in the synthetic (imputed) data file, or at least when these two distributions are "very close". The discrepancy between the joint distribution of the variables of interest (a) in the population, and (b) in the synthetic data file is usually referred to as matching noise Paass (1986). Attempts at evaluating the "closeness" of the empirical distribution of imputed data to the empirical distribution of "real" data have been performed in the literature, see D'Orazio et al. (2006). In a nonparametric setting an important role is played by hot-deck methods, as well as k-nearest neighbor (kNN) methods. Their properties are studied in Marella et al. (2008), where both theoretical and simulation results are obtained.

As a matter of fact, the CIA is usually a misspecified assumption, and external auxiliary information is most of the times not available. The lack of joint information on the variables of interest is the cause of uncertainty on the model of (X, Y, Z). The problem is that sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X, Y, Z). In other terms, the adopted statistical model is not identifiable on the basis of sample data. Hence, a third group of techniques that does not directly aim at reconstructing a complete data set is introduced. This group of techniques addresses the so-called identification problem. The main consequence of the lack of identifiability is that some parameters of the model cannot be estimated on the basis of the available sample information. Instead of point estimates, one can only reasonably construct sets of "possible

point estimates", compatible with what can be estimated (i.e., each point estimate is obtained by imposing a model which is compatible with the estimable distributions Y|X and Z|X).

These sets (usually intervals) formally provide a representation of uncertainty about the model parameters (note that these intervals are not confidence intervals, the problem is not sampling variability, but the lack of joint information on Y and Z).

In this setting, the main task consists in constructing a coherent measure that can reasonably quantify the uncertainty about the (estimated) model. From an operational point of view, a measure of uncertainty essentially quantifies how "large" is the class of models estimated on the basis of the available sample information. The smaller the measure of uncertainty, the smaller the class of estimated models. Preliminary studies on this have been considered in Kadane (1979), Rubin (1986), Raessler (2002), D'Orazio et al (2006, Chapter 4). A thorough discussion on uncertainty measures is in Conti et al (2012).

When dealing with samples drawn according to complex survey designs, there is the problem of how to use the possibly different survey weights in a statistical matching context. Up to now there are essentially two distinct approaches.

1. File concatenation. This approach was suggested by Rubin (1986) and consists in defining the probabilities of inclusion that the units in the A sample would have had if the survey design of sample B was adopted (say π_a^B, a=1,…n_A), and the probabilities of inclusion that the units in the B samples would have had if the survey design of sample A was adopted (say π_b^A, b=1,…n_B). Then, the file obtained concatenating the two samples will have n_A+n_B units with probability of inclusion: $\pi_h^{A\cup B} = \pi_h^A + \pi_h^B - \pi_h^{A\cap B}$, h=1,…, n_A+n_B, where the last term indicates the probability of inclusion of a unit in the intersection between the two samples. Most of the times this last probability is negligible, and as suggested by Rubin it can be eliminated in the formula. This is not the case when, for instance, there are "take-all" strata in the two samples with a non-empty intersection (as it is typical for enterprise surveys, where take-all strata usually consist of large enterprises). Rubin suggests to use multiple imputation in order to fill in the missing data in the concatenated file.
2. Calibration. This approach was suggested by Renssen (1998), and consists in estimating all the distributions of X, Y|X and Z|X from A and B after a calibration step that makes the two surveys coherent on the common information (X). These distributions allow to apply statistical matching procedures under the CIA (Renssen suggests to use imputation by regression functions). Renssen studies also the case a complete third sample C is available and suggests two different procedures for making information on A, B and C coherent by means of calibration procedures. This use of an external auxiliary file C allows to avoid the assumption of conditional independence for Y and Z given X. Again, a complete file can be obtained by using imputation by regression.

3. Design issues

This section has been taken from the WP2 of the ESSnet on ISAD (integration of surveys and administrative data), Section 3.1 (Scanu, 2008a).

Figure 1 represents the steps that need to be performed for solving a statistical matching problem.

1) A key role is represented by the choice of the target variables, i.e., of the variables observed distinctly in two sample surveys. The objective of the study will be to obtain joint information on these variables. This task is important because it influences all the subsequent steps. In particular, the matching variables (i.e., those variables used for linking the two sample surveys) will be chosen according to their capacity to preserve the direct relationship between the target variables.

2) The second step is the identification of all the common variables in the two sources (potentially all these variables can be used as matching variables). Not all these variables can actually be used. The reasons can be different, as lack of harmonisation between the variables. To this purpose, some steps need to be performed as the harmonisation of their definition and classification, the need to take only accurate variables whose statistical content is homogeneous.

3) Once the common variables have been cleaned of those variables that cannot be harmonised, it is necessary to choose only those that are able to predict the target variables. To this purpose, it is possible to apply some statistical methods whose aim is to discover the relationship between variables, as statistical tests or appropriate models.

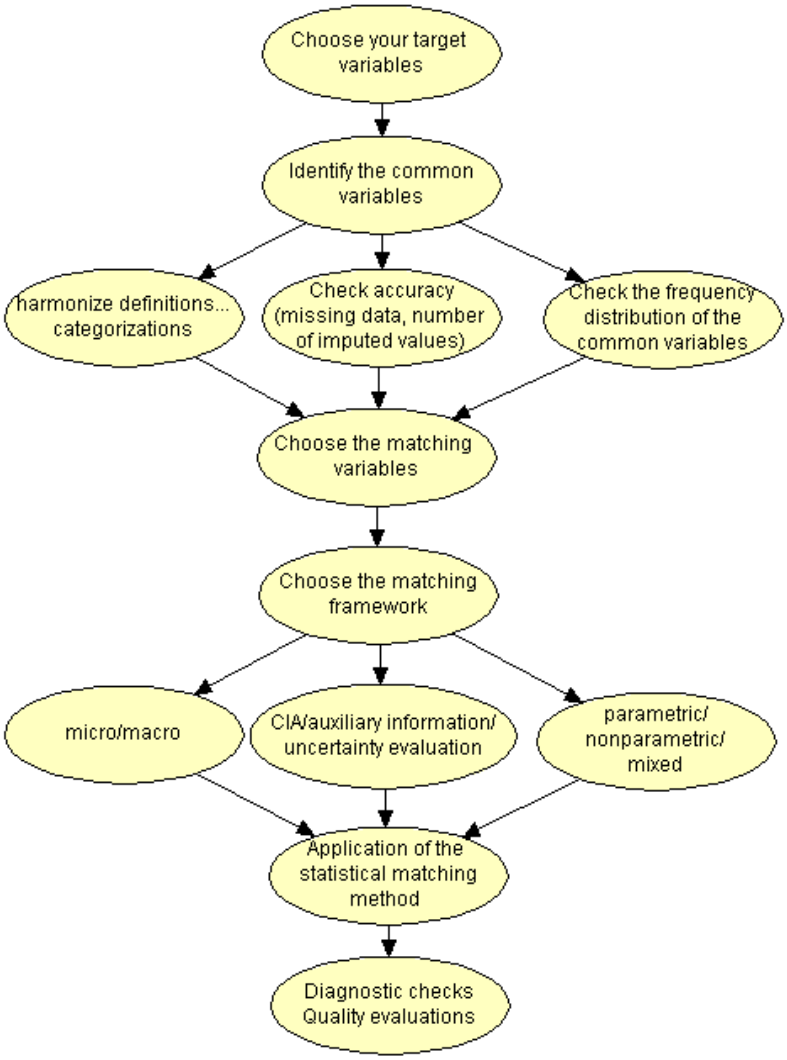

Figure 1: workflow of the actions to perform in statistical matching

4) As already introduced in the beginning, the statistical matching aim can be solved in different ways:

   a. By a micro objective (i.e., construction of a complete data file with joint information on X, Y, and Z) or a macro objective (i.e., estimation of a parameter on the joint distribution of (Y,Z), (Y,Z|X), (X,Y,Z))

   b. By the use of specific models (as the conditional independence assumption), the use of auxiliary information, or the study of uncertainty

   c. By parametric, nonparametric or mixed procedures (this will be specified in Section "Statistical matching methods").

5) Once a decision has been taken, the procedure is applied on the available data sets.

6) Quality evaluations of the results are the final step to perform.

Chapter 3 of the Report on WP2 of the ESSnet on ISAD describes in detail all the previous steps. The previous steps correspond to choices taken by the researcher that is performing a statistical matching application. What happens if some of the steps cannot be performed? This problem is especially connected with step 3, i.e., on the choice of the matching variables. If the common variables are unable to predict the target variables (e.g., they are independent of the target variables), statistical matching cannot be performed, because the common variables do not add any information on the relationship between the target variables.

4. Available software tools

The ESSnet on Integration of Surveys and Administrative data (ISAD) dealt with the problem of software tools in data integration. Workpackage 3 includes a thorough discussion on the available software tools (see Chapter 2, Scanu 2008b).

SAMWIN (Sacco, 2008): The software package SAMWIN was built for the production of an integrated archive for the social accounting matrix. This integrated archive was built by means of statistical matching techniques based on nonparametric imputation methods (hot-deck). For this reason, SAMWIN includes only matching algorithms based on the donors, more precisely distance hot-deck algorithms. The platform for SAMWIN is Visual Studio 6 (Visual C++). The developer is Giuseppe Sacco. Any question on SAMWIN should be sent to the email address sacco@istat.it.

StatMatch (D'Orazio, 2011). This is an R package consisting of functions for the implementation of statistical matching methods based on imputation procedures, under both the conditional independence assumption and the use of auxiliary information It also includes functions for the evaluation of uncertainty

SPlus codes (Raessler, 2002). These codes were written by Raessler for the implementation of proper multiple imputation methods for statistical matching in a Bayesian context.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Conti, P. L., Marella, D., and Scanu, M. (2012), Uncertainty analysis in statistical matching. *Journal of Official Statistics* **28**, 1–21.

D'Orazio, M. (2011), *Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment*. Vignette for the application of the R package StatMatch, available on CRAN and at http://www.cros-portal.eu/content/wp3-development-common-software-tools.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical matching: theory and practice*. Wiley, Chichester.

Kadane, J. B. (1978), Some Statistical Problems in Merging Data Files. Compendium of Tax Research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179. (Reprinted in 2001, *Journal of Official Statistics* **17**, 423–433).

Marella, D., Scanu, M., and Conti, P. L. (2008), On the Matching Noise of Some Nonparametric Imputation Procedures. *Statistics and Probability Letters* **78**, 1593–1600.

Okner, B. A. (1972), Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File. *Annals of Economic and Social Measurement* **1**, 325–362.

Paass, G. (1986), Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. In: G.H. Orcutt and H. Quinke (eds.), *Microanalytic Simulation Models to Support Social and Financial Policy*, Elsevier, Amsterdam.

Raessler, S. (2002), *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Lecture Notes in Statistics, Springer Verlag, New York.

Renssen, R. H. (1998), Use of Statistical Matching Techniques in Calibration Estimation. *Survey Methodology* **24**, 171–183.

Ridder, G. and Moffitt, R. (2007), The Econometrics of Data Combination. In: J. J. Heckmann and E. E. Leamer (eds.), *Handbook of Econometrics*, vol. 6A, Elsevier, Amsterdam.

Rodgers, W. L. (1984), An Evaluation of Statistical Matching. *Journal of Business and Economic Statistics* **2**, 91–102.

Rubin, D. B. (1986), Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* **4**, 87–94.

Sacco, G. (2008), SAMWIN: a software for statistical matching. Document of WP3 of the *ESSnet on Integration of Surveys and Administrative Data*, available at http://cenex-isad.istat.it/archivio/Technical_reports_and_documentation/software_on_statistical_matching/SAMWIN_manual.pdf.

Scanu, M. (2008a), The practical aspects to be considered for statistical matching. Section 3.1 of the *Report on WP2 of the ESSnet on Integration of Surveys and Administrative Data* (http://www.cros-portal.eu/content/isad-finished)

Scanu, M. (2008b), Software tools for statistical matching. Chapter 2 of the *Report on WP3 of the ESSnet on Integration of Survey and Administrative Data* (http://www.cros-portal.eu/content/isad-finished).

Sims, C. A. (1972), Comments and Rejoinder (On Okner (1972)). *Annals of Economic and Social Measurement* **1**, 343–345 and 355–357.

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**, 59–79.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Donor Imputation
3. Weighting and Estimation – Main Module
4. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

1. StatMatch (R package)
2. SamWin

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.1: Integrate data

Administrative section

14. Module code

Micro-Fusion-T-Statistical Matching

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2012	first version	Mauro Scanu	Istat (Italy)
0.2	02-05-2012	second version	Mauro Scanu	Istat (Italy)
0.3	25-09-2013	EB comments	Mauro Scanu	Istat (Italy)
0.3.1	03-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:59