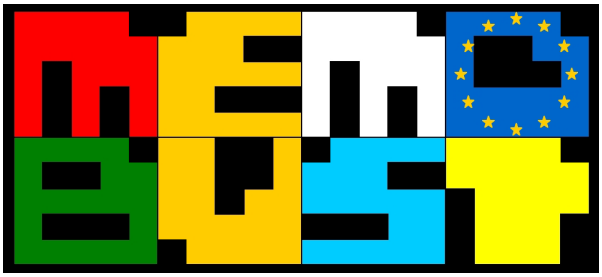

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Data Fusion at Micro Level

Contents

General section

1. Summary

This module gives a general overview of problems and methods concerning the integration of several data sources for statistical purposes. It is focused on the case of integration at micro level, which means to integrate data sources composed of units (input: micro) in order to obtain still a data set composed of units (output: micro).

2. General description

There are more statistical data produced in today's society than ever before. These data are analysed and cross-referenced for innumerable reasons. In the case of National Statistical Institutes (NSIs) the joint analysis of two or more statistical and administrative sources is a result of a rational organisation of all available informative sources and, among all, it allows the reduction of survey costs, the response burden and to enrich the information already held on such units by means of adding new data from other sources enabling for instance the analysis of relationships among variables observed in different data sources. Nevertheless, the integration process must deal with many different problems.

This module gives a general overview of problems and methods concerning the integration of several data sources for statistical purposes. Integration is made at micro level, which means the integration of data sources composed of units (input: micro) with the aim of obtaining still a data set composed of units (output: micro).

The problems discussed in this section essentially deal with two questions: how to fuse different data sources and how to manage consistency problems.

The section is mainly based on the documents produced in the two European projects funded by Eurostat on data integration: the ESSnet *Integration of Surveys and Administrative Data* carried out during 2006-2007 (ISAD, 2006), and the Essnet on *Data Integration* during 2010-2011 (DI, 2011).

2.1 Data fusion

An important element to take into account when different data sets are fused concerns if they are composed of

1) (almost) the same units;

2) different units.

The first case is typical of integration between registers and sample surveys, while the second typically happens when integration is related to sample surveys.

This distinction is important since different methods are required. Essentially, in the first case we resort to statistical classification methods and in this context they are referred to as record linkage procedures, while in the second we mainly resort to imputation methods that are usually referred to as statistical matching techniques.

2.1.1 Record linkage/Object Matching

Record linkage (also known as *object matching*) consists in identifying pairs of records coming from different data sets, which belong to the same entity, on the basis of the agreement between common

variables (name, address, telephone,...). It may happen that the same units have different values for the common variables in the two data sets, for instance because of a change in the telephone number or because some error affects data (Herzog et al. 2007).

In general, key variables are compared in order to understand whether a pair of observations from the two files is either a match or an unmatch.

The results of the comparisons may be used in different ways resulting in different record linkage/object matching methods that can be classified as:

1) deterministic approaches;

2) probabilistic approaches.

Deterministic approaches are characterised by the use of formal decision rules. In this framework, some algorithms are developed for linking data, for details see the modules "Micro-Fusion – Unweighted Matching of Object Characteristics" and "Micro-Fusion – Weighted Matching of Object Characteristics".

Probabilistic approaches make an explicit use of probabilities for deciding when a given pair of records is actually a match given the results of the comparison of the key variables. The probabilities allow to quantify the degree of uncertainty in a match/unmatched pair of observations and may help the researcher to take decisions within a formalised probabilistic setting allowing in some cases the estimation of errors associated to the performed action.

The procedure proposed by Fellegi and Sunter (1969) is one of the reference techniques for probabilistic record linkage. They deal with the problem of linkage by using a latent model, where the latent variable describes the two populations of matches and unmatches. For each pair of observations, the probabilities of belonging to the two populations are computed according to the values obtained by the comparison of the key variables. Intermediate situations where the pairs cannot be classified with a high probability in one of the two populations may arise. For these units a clerical review is needed. The advantage of using a probabilistic approach is that it is allowed to estimate the errors associated to the decision taken in the classification step, and they can be used to establish a proper methodological framework for a statistical procedure to decide links, or can be used when assessing the quality of estimates obtained with integrated data that are affected by a further source of uncertainty due to the linkage process. The latter is a problem that still requires further investigations, for details see Di Consiglio and Tuoto (2013) and references therein.

The probability distributions of the results of the comparisons of the key variables for respectively match and nonmatch populations are essential elements of the probabilistic record linkage approach. They are generally not known and an estimation step is needed.

It is worthwhile to remark a peculiarity of the linkage procedures as previously formalised. The number of matches is naturally much less than the number of unmatches. Without loss of generality, let n_A be the number of observations in the first file A and n_B the number of observations in the second file B to be linked, where $n_A \leq n_B$. In the most favorable situation there are n_A matches out of the $n_A \times n_B$ pairs of units. A very low proportion of matches with respect to all the pairs of units may result in a not reliable estimation result because the unmatches tend to overwhelm the information coming from the rare population of matches. This problem is alleviated by using blocking procedures, that is to split records into groups (blocks) and comparing only the units belonging to the same group. When data

sets are large, this task is also particularly important from an operational point of view since it can be computationally unfeasible to make a large number of comparisons. On the other hand restricting the matches within the block may be dangerous because of the exclusion of some possible matches. Suggestions for choosing the blocking procedures can be found in ISAD (2006).

For more details on probabilistic record linkage and the Fellegi-Sunter procedure see the modules "Micro-Fusion – Probabilistic Record Linkage" and "Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage".

2.1.2 Statistical matching

Statistical matching (also named *data fusion* or *synthetic matching*) refers to a series of methods whose objective is the integration of two (or more) data sources referring to the same target population. The data sources are characterised by the fact they all share a subset of variables (common variables) and, at the same time, each source observes distinctly other subsets of variables. Moreover, there is a negligible chance that data in different sources observe the same units (disjoint sets of units).

In the simplest case of two samples (data sources A and B), the classical statistical matching framework is represented in Figure 1:

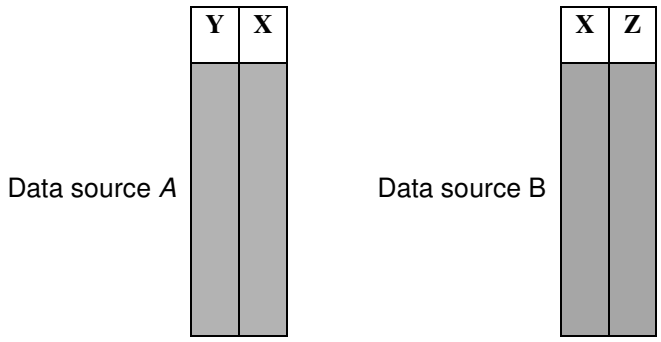

Figure 1. The statistical matching setting

The common variables are denoted by X, the set of variables Y are observed only in A but not in B, and Z are observed in B but not in A (Y and Z are not jointly observed).

Statistical matching methods aim at integrating the two sources in order to study the relationship existing among the two sets of variables not jointly observed, i.e., Y and Z or, more in general, to study how X, Y and Z are related.

In the *micro approach*, the statistical matching objective is the construction of a complete "synthetic" file, that is a file where X, Y and Z are jointly present. The term synthetic refers to the fact that this file is not the result of a direct observation of all the variables on a set of units belonging to the population of interest, but it is obtained exploiting information in the observed distinct files. For example, in the case of the data sets as in the previous figure, a synthetic file is the one in Figure 2, where the file A is filled in with the synthetic values $\tilde{\mathbf{Z}}$ by exploiting the joint information regarding X and Z observed in B.

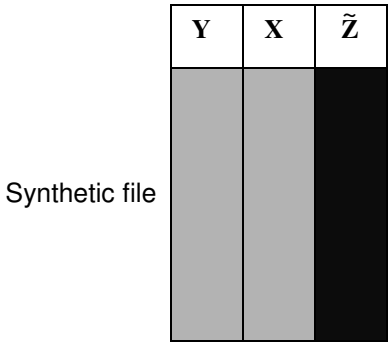

Figure 2. Statistical matching at micro level

The file is generally obtained through imputation techniques that may resort to parametric models, nonparametric models, and a mixture of them. For parametric imputations an important role is assigned to regression models, in this case the regression of Z (dependent variable) vs the covariate X is estimated on B. The prediction $\tilde{\mathbf{Z}}$ is obtained by applying the estimated model to the X values in A. The most frequently used nonparametric models refer to the family of hot-deck imputation methods. By considering the situation in Figure 2 the value $\tilde{\mathbf{Z}}$ for a given unit is obtained by taking the *Z* value observed in the most similar observation in the data set B, where the similarity is computed with respect to the X values. Details of those imputation methods are given in the module "Micro-Fusion – Statistical Matching Methods". In this framework, whatever matching procedure is used, the results will be based on the conditional independence assumption (CIA) of *Y* and *Z* given *X*, which means that the results can be considered acceptable only if, roughly speaking, the information in X is so rich that it can explain the relationships between Y and Z. Another way of looking at the CIA is that the probability distribution of Y conditionally on Z and X depends only on X. Broadly speaking, it is not important to look at the value of Z to be imputed to the observation having Y, once you know its X value. For instance, let us suppose that the variable Y denotes the income, let Z be the level of consumption and let X represent the geographical area. The CIA states that, for example, in order to impute a consumption level to a certain unit, it is not important to know the level of income once you know the geographical area where the unit belongs to.

CIA is a strong assumption that cannot be always assumed and unfortunately it cannot be tested with the available data since Y and Z are not jointly observed. In order to avoid the CIA the use of auxiliary information may be useful and, for this reason in any statistical matching process, enough time should be devoted to search for any kind of additional information. In order to avoid CIA the auxiliary information should be about the variables not jointly observed. It can be in the form of a third file where either (X,Y,Z) or (Y,Z) are jointly observed (e.g., outdated data), or as plausible values of the inestimable parameters of either (Y,Z|X), or (Y,Z). Information is useful also when it is not exactly about the Y, Z and X but it refers to proxy variables (e.g., outdated data where numerical variables are observed as categorical/ordinal by collecting only ranges). This kind of information is generally sufficient to determine a model without assuming the CIA. Specific methods involving the use of auxiliary information are described in D'Orazio et al. (2006).

Another kind of auxiliary information is that provided by logical rules relating the variables Y and Z (usually named edit rules in the editing procedures, see the topic "Statistical Data Editing"). One

example of logical rule is that it is generally not acceptable that a ten-year-old person is married. This kind of auxiliary information is not generally sufficient to determine a unique model alternative to the CIA, however it can be useful to restrict the possible statistical models compatible with the data at hand. Their use is important to increase the quality of the matching procedure.

For more details on statistical matching see the modules "Micro-Fusion – Statistical Matching" and "Micro-Fusion – Statistical Matching Methods".

2.1.3 Micro-integration

A problem that must be dealt with when integration of different data sources is performed is that of consistency. Procedures in order to make coherent and consistent data at micro level are generally unavoidable. The set of tasks with this purpose are named micro-integration (see Bakker 2011).

A definition of micro-integration is given in Bakker (2011): "*Micro-integration is the method that aims at improving the data quality in combined sources by searching and correcting for the errors on unit level.*"

The term "error" should be understood in a broad sense, Bakker refers to measurement and representation errors.

Representation errors exist if the target population is incompletely described by the data (e.g., over-coverage, under-coverage, ...).

Measurement errors exist if characteristics of the population elements are not correctly described. These errors may have different causes. By using information from different sources, these errors can be detected and corrected. *Harmonisation* is the correction on a conceptual level (for instance harmonising the definition of the variables), while for the correction on data level Bakker uses correction for measurement errors (also known as data reconciliation).

Representation and harmonisation problems are dealt with case by case, there are no general algorithms for this purpose. Since the focus of the module is on general statistical techniques, we do not discuss the so far mentioned problem, the interested reader may refer to Van der Laan (2000).

As far as data reconciliation is concerned, some techniques can be applied. In the module "Micro-Fusion – Reconciling Conflicting Microdata" it is discussed the problems arising when linked records do not satisfy edit-rules and an adjustment step is necessary to integrate the different pieces of information, (the data sources and the edit-constraints), to obtain consistent integrated microdata. One possible strategy is to adjust data in order to satisfy edit-rules. The module "Micro-Fusion – Prorating" describes a ratio adjustment method for balance edits. It solves the possible inconsistencies for each constraint separately by distributing the differences between the total and the items composing the total. The main advantages are that is easy to interpret and to apply.

More refined methods are introduced in literature. In the module "Micro-Fusion – Minimum Adjustment Methods" the data reconciliation task is formalised as a constrained minimisation problem, that is to find the final imputed values such that 1) they differ as little as possible from the observed data and such that 2) they satisfy the edit-rules. The evaluation of changes are computed according to different distances: least squares, weighted least squares and Kullback-Leibler. The procedure is presented according to two different settings: 1) one data set is considered more reliable

than the other, 2) data sets to be integrated are considered equally reliable. The constraints considered in the minimisation problem are linear.

The method described in the module "Micro-Fusion – Generalised Ratio Adjustments" aims to make the adjustments as uniform as possible, and in contrary to the other methods, the method can result in adjustments to variables that are not involved in the constraints. This may be useful to preserve relations between variables that are not connected by edit rules.

Procedures that aim at reaching consistency at output level by modifying data at micro level are still classified as micro-integration procedures. According to this definition, a problem that frequently arises is that of consistency of published figures (for instance frequency tables) when a survey is enriched with register data. An example can be useful to understand the problem. Let us suppose we have a situation like that depicted in Figure 3,

Units	$X_1,...,X_j$	$Y_1,...,Y_k$	w
1 . n			
.			
N			

Figure 3. Micro-integration of a register and a survey.

where grey cells represents the available information, variables $X=(X_1,...,X_j)$ are the variables from the register and are observed for all the units in the population composed of N units, variables $Y=(y_1,..,y_k)$ are the variables collected in the survey and observed only on a subset of units of the population (sample size n) and finally $w=(w_1,...,w_n)$ are the sampling weights associated to each sample unit. Estimates for the parameters (totals, frequency tables) of the variables X can be obtained by using the sampling weights w. The estimates should be consistent with the value computed according to the registered data. This is usually accomplished by using calibration procedures, i.e., by changing as little as possible the values of w such that the two parameters are the same. We notice that also in this case there is a change at micro level (sampling weights) in order to have consistent outputs. The problem becomes more difficult when some of the variables X in the register are not used in the calibration procedure because for instance the treatment of a high number of variables in the calibration is not feasible. In this case an iterative procedure named *consistent repeated weighting* is proposed by (Houbiers et al., 2003; Kroese et al., 2000; Renssen et al., 2001; Houbiers, 2004). It is based on the repeated application of the regression estimator and generates a new set of weights for each table that is estimated. All the tables considered will be consistent.

An alternative approach to reach consistency of the estimates in the latter situation is named *mass imputation*. It consists in imputing all the variables $y_1,..,y_k$ in order to obtain a final rectangular data set. Whitridge and Kovar (1990) discuss the practical advantages of such a procedure, in fact the

estimates obtained with such a completed data set are naturally consistent, however Kroese et al., (2000) warn about the fact of having in practice enough degrees of freedom to get a model for imputing data such that all the possible relationships are obtained.

3. Design issues

4. Available software tools

Workpackage 3 of the Essnet on Data Integration includes a thorough discussion on the available software tools (see Scanu, 2008b, Chapter 2).

Some specific references are reported in the following.

StatMatch is an R-package for statistical matching (D'Orazio, 2011) freely available on the website http://cran.r-project.org/.

Relais is a freely available software developed by Istat for probabilistic record linkage, downloadable from http://www.istat.it/it/strumenti/metodi-e-software/software/relais.

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

Bakker, B. F. M. (2011), Micro-Integration: State of the art. In: *Report WP1: State-of-the-art on Statistical Methodologies for Data Integration*, ESSNET on Data Integration, available at http://www.cros-portal.eu/content/wp1-state-art.

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*. John Wiley, Chichester.

Di Consiglio, L. and Tuoto, T. (2013), Challenges in estimation on probabilistically linked data. Proceedings of NTTS 2013, available at http://www.cros-portal.eu/sites/default/files//NTTS2013fullPaper_112.pdf.

DI (2011), Essnet on Data Integration, http://www.cros-portal.eu/content/data-integration-1.

Fellegi, I. P. and Sunter, A.B. (1969), A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Data Quality and Record Linkage Techniques*. Springer.

Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H., and Snijders, V. (2003), Estimating consistent table sets: position paper on repeated weighting. Discussion paper 03005, Statistics Netherlands, Voorburg.

Houbiers, M. (2004), Towards a Social Statistical Database and unified estimates at Statistics Netherlands. *Journal of Official Statistics* **20**, 55–75.

ISAD (2006), ESSnet on Integration of Survey and Administrative Data, http://www.cros-portal.eu/content/isad-finished.

Kroese, A. H., Renssen, R. H., and Trijssenaar, M. (2000), Weighting or imputation: constructing a consistent set of estimates based on data from different sources. In: P. G. Al and B. F. M. Bakker (eds.), Special Issue: Re-engeneering social statistics by micro-integration of different sources, *Netherlands Official Statistics* **15**, Summer, 23–31.

Pannekoek, J. (2011), Models and algorithms for micro-integration. In: *Report on WP2: Methodological developments*, ESSNET on Data Integration, available at http://www.cros-portal.eu/content/wp2-development-methods.

Renssen, R. H., Kroese, A. H., and Willeboordse, A. (2001), Aligning estimates by repeated weighting. Research paper 491-01-TMO, Statistics Netherlands, Voorburg/Heerlen.

Van der Laan, P. (2000), Integrating Administrative Registers and Household Surveys. In: P. G. Al and B. F. M. Bakker (eds.), Special Issue: Re-engeneering social statistics by micro-integration of different sources, *Netherlands Official Statistics* **15**, Summer, 7–15.

Whitridge, P. and Kovar, J. G. (1990), Use of mass imputation to estimate for subsample variables. In: *Proc. Bus. Econ. Statist. Sect.*, American Statistical Association, Washington, DC, 132–137.

Interconnections with other modules

8. Related themes described in other modules

1. Micro-Fusion – Object Matching (Record Linkage)
2. Micro-Fusion – Probabilistic Record Linkage
3. Micro-Fusion – Statistical Matching
4. Statistical Data Editing – Main Module
5. Imputation – Main Module
6. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1. Micro-Fusion – Unweighted Matching of Object Characteristics
2. Micro-Fusion – Weighted Matching of Object Characteristics
3. Micro-Fusion – Fellegi-Sunter and Jaro Approach to Record Linkage
4. Micro-Fusion – Statistical Matching Methods
5. Micro-Fusion – Reconciling Conflicting Microdata
6. Micro-Fusion – Prorating
7. Micro-Fusion – Minimum Adjustment Methods
8. Micro-Fusion – Generalised Ratio Adjustments

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.

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-Data Fusion at Micro Level

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2012	first version	Marco Di Zio	Istat (Italy)
0.2	06-04-2012	second version	Marco Di Zio	Istat (Italy)
0.3	11-11-2013	revision based on EB comments for preliminary release. Final template is used	Marco Di Zio	Istat (Italy)
0.3.1	18-11-2013	preliminary release		
0.4	19-12-2013	final release based on EB comments	Marco Di Zio	Istat (Italy)
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 17:56