# Integration and imputation of survey data in R: the StatMatch package

**Article** *in* Revista română de statistică: organ al Comisiei Naționale pentru Statistică · June 2015

CITATIONS
13

READS
417

**1 author:**

Marcello D'Orazio
Italian National Institute of Statistics
**56** PUBLICATIONS **481** CITATIONS

SEE PROFILE

# Integration and imputation of survey data in R: the StatMatch package

**Marcello D'ORAZIO** (madorazi@istat.it)
Italian National Institute of Statistics (Istat)

## ABSTRACT

*Statistical matching methods permit to integrate two or more data sources with the purpose of investigating the relationship between variables not jointly observed. Recently these methods received much attention as valid alternative to produce new statistical outputs.*

*The paper provides an overview on the statistical matching methods implemented in the package StatMatch for the R environment, focusing on the most widespread methods and how they were improved. Particular attention is devoted to hot deck matching methods, strictly related to the ones developed for the imputation of missing values. The corresponding functions in StatMatch are very powerful and are flexible enough to be applied for imputing missing values in a survey. The paper tackles also the problem of matching data from complex sample surveys, a very important topic in National Statistical Institutes. Finally it is described the concept of uncertainty characterizing the statistical matching framework and how this alternative approach can be exploited for different purposes.*

**Keywords***: statistical matching, hot deck imputation methods, uncertainty*

## INTRODUCTION

The increasing demand for new statistical outputs poses several challenges for National Statistical Institutes; setting up new surveys or modifying existing ones is often unfeasible due to budget constraints and to avoid the increase of respondents' burden. In this framework, a better exploitation of the existing data sources (survey data, administrative registers, etc.) becomes a key topic. In particular the integration of the available data sources represents a viable solution to satisfy the demand for new statistics. A typical example is represented by the statistics concerning the people living conditions; in most of the countries producing such information requires the availability of data about the consumption as well as on the income, but these data are collected in distinct surveys and it is difficult to design and carry out a unique survey to investigate both the phenomena.

In statistics, the integration techniques can be divided in two groups: (i) *record linkage* and (ii) *statistical matching*. Record linkage techniques aim at identifying pairs of records, in different data sources, which refer to the same entity

(person, household, enterprise, etc.). *Statistical matching* (or *data fusion*) techniques are meant at investigating the relationship of variables not jointly observed in a single data source. The matching techniques, encompass a wide set of statistical methods ranging from imputation of missing values to methods for estimating parameters in the presence of missing values. The techniques do not necessarily involve the derivation of an integrated ("fused") data source; the estimation of parameters (e.g. correlation coefficient) be achieved by applying appropriate statistical methods, without integrating the data sources at micro level.

Statistical matching methods include most of the of the methods developed to impute the missing values in a survey (donor based, regression imputation, etc.). This may facilitate application of matching, because, in theory, it is possible to resort to software packages developed for imputation purposes; unfortunately, in practice, it may not be simple to adapt these software packages to the matching framework. The lack of software packages for performing statistical matching led to the development of the package **StatMatch** (D'Orazio, 2015), an open source additional package for the R environment (R Core Team, 2014).

This paper provides an overview of the statistical matching methods implemented in **StatMatch**, with a major emphasis on the recent work done to enhance some of the "traditional" matching methods.

## STATISTICAL MATCHING

The "basic" statistical matching (SM henceforth) framework consists of two data sources, $A$ and $B$, sharing a set of variables $X$ (common variables) while the variable $Y$ is available just in $A$ and the variables $Z$ is available only in $B$. In practice $Y$ and $Z$ are not jointly observed and the goal of SM consists in exploring their relationship.

Several SM methods have been proposed (for major details see D'Orazio *et al.*, 2006b). A broad classification of the methods can be derived by considering their goal and the approach followed. When the objective is the creation an integrated (*fused*) data set it is used the term *micro*; on the contrary, the term *macro* refers to cases where the target is the estimation of a set of parameters, such as correlation/ regression coefficients, frequencies in contingency tables, etc. The approach to SM can be: (i) *parametric*, when it is explicitly considered a statistical model to describe the relationship between $X$, $Y$ and $Z$; (ii) *nonparametric*, when the model is not assumed; and, (iii) *mixed* when there is a mix of parametric and nonparametric methods. The Table 1 provides a summary overview of some SM methods according to the objective and to the approach.

**Methods for Statistical matching according to approaches and**

*Table 1*

| Objective of SM | Parametric | Nonparametric | Mixed |
|---|---|---|---|
| Macro | - methods for estimation of model parameters in the presence of missing values | - estimation of the empirical cumulative distribution<br>- kernel density estimators | |
| Micro | - conditional mean matching<br>- stochastic regression imputation<br>- … | - hot deck imputation procedures | - combination of predictive mean matching and hot deck imputation<br>- … |

The choice of the SM method depends on: (1) the target of the matching, an integrated data set (micro case) or estimates of parameters (macro case); (2) the framework for inference, traditional *model based* or *design based* (this latter one is typical in finite population survey sampling theory); and (3) the underlying assumptions.

It is important to stress that most of SM methods proposed in literature relies upon a very strong untestable assumption. In practice, all the methods which use just the common information $X$ to match the data sources, are implicitly assuming that $Y$ and $Z$ are independent once conditioning on $X$:

$$f(x,y,z) = f(y|x) \times f(z|x) \times f(x) \qquad [1]$$

This *conditional independence* (CI) assumption cannot be tested in the "basic" SM setting and, unfortunately, it is seldom valid in real world applications. It can be bypassed in presence of some auxiliary information concerning the relationship between $Y$ and $Z$; e.g. a previous estimate of the parameters of interest or, better, a third additional data source in which $Y$ and $Z$ are jointly observed. In alternative it is possible to tackle the SM application by focusing on the *uncertainty* characterizing the target parameter to estimate (macro objective); the uncertainty is due to the fact that the interest variables, $Y$ and $Z$, are not jointly observed. Methods for the exploration of uncertainty in the case of categorical variables will be presented in Section 4.

Most of the SM methods (typically the macro ones) assume that the $A$ and $B$ consist of independent and identically distributed (i.i.d.) observations. Unfortunately, in National Statistical Offices data sources often originate from sample surveys carried out to investigate phenomena in given finite populations. In such cases the samples are selected by means of complex sampling designs (involving stratification, clustering, etc.) and therefore it is difficult to maintain the i.i.d. assumption; it would

mean that the sampling design can be ignored (see Särndal *et al.*, 1992, Section 13.6). In literature there are relatively few SM methods to deal with such data, in particular the Rubin's *file concatenation* approach (micro) and Renssen's *weights' calibration* based approach (cf. D'Orazio, 2011a). This latter method will be shortly described in Section 3.

***Statistical matching in R: the package StatMatch***

The lack of software for performing SM led to the development of the package **StatMatch** for the R environment. The decision to work in R is due to its characteristics: a free open sources environment which easily permits to implement new statistical methods, offering the chance of re-using functions developed by researchers in Academia or other national statistical Institutes.

The first appearance of **StatMatch** on the CRAN (Comprehensive R Archive Network) dates back to 2008; it was mainly based on R codes provided in the Appendix of D'Orazio *et al.* (2006b). Since then a number of updates have been released. A valuable contribution to the improvement of **StatMatch** came from the Eurostat's ESSnet project on "Data Integration" (D'Orazio 2011b). The latest version of **StatMatch** (v. 1.2.3) was released on January 2015.

The functions in **StatMatch** can be divided in five main groups:

- functions to perform nonparametric SM at micro level via *hot deck imputation* methods (NND.hotdeck, RANDwNND.hotdeck, rankNND.hotdeck);
- a function to perform *mixed* SM at macro or micro level in presence of continuous variables being distributed according to the multivariate normal distribution (mixed.mtc);
- functions to integrate data from complex sample surveys through calibration of weights as proposed by Renssen (1998) (harmonize.x and comb.samples);
- functions to explore uncertainty on the cells of a contingency table when all the variables $X$, $Y$ and $Z$ are categorical (Frechet.bounds.cat and Fbwidhts.by.x);
- other functions to perform tasks before or after the matching step, e.g. to select the matching variables; to compute distances etc.

## HOT DECK STATISTICAL MATCHING METHODS

The SM methods based on hot deck imputation procedures are frequently encountered in SM applications. These methods are simple and easy to apply being derived directly from the corresponding methods developed to impute missing values in a survey (Singh *et al*, 1993). They do not need the specification of a model and the subsequent estimation of parameters. They are applied mostly when the goal is micro, given that they are designed to provide a fused or "synthetic" data set (synthetic because it is not obtained as a direct output of data collection). The application of hot

deck methods in the matching framework requires the identification of the *recipient* data source, while the other one becomes the *donor*: it donates to the recipient the values of the variable initially missing. The recipient should have fewer observations than the donor to avoid typical drawbacks of this kind of procedures (multiple usage of donors, risk of biased marginal distribution of the imputed variables, etc.).

Methods commonly used are (Singh *et al*, 1993): the *nearest neighbor distance hot deck*; the *random hot deck*; and the *rank hot deck*.

***Nearest neighbor distance hot deck***

Nearest neighbor distance hot deck is very common; the method searches the closest donor for each recipient unit, in terms of distance computed on the chosen subsets $X_M$ of the common variables $X$ ( $X_M \subseteq X$ ); the $X_M$ are called *matching variables*. In presence of many recipient and donor records the effort for computing the distance matrix may become non-negligible; for this reason it is suggested to limit the search of the donors in proper subgroups of units sharing the same characteristics, said *donation classes*; e.g. for a recipient with gender male the search of the donor will be limited just to the males.

When applying the nearest neighbor distance hot deck a critical issue concerns the choice of the matching variables and of the distance functions. Obviously the distance function depends on the nature of variables; for instance in the presence of mixed types of variables (categorical and continuous) it is possible to resort to the Gower's dissimilarity which avoids the transformation of the variables (substitution on the categorical variables with the corresponding dummies).

Nearest neighbor distance hot deck is implemented in the function NND. hotdeck of **StatMatch**. It supplies a variety of well-known distance functions (Manhattan, Euclidean, etc.) including the Gower's dissimilarity. Recently the *maximum distance* ( $L^{\infty}$ norm) has been added; this function works on the true observed values (continuous variables) or on transformed ranked values, following the suggestion by Kovar *et al.* (1988); the transformation (ranks divided by the number of units) removes the effect of different scales and the new values are uniformly distributed in the interval [0,1].

As far as the choice of the matching variables is concerned, there is a wide series of methods that can be applied, for instance the ones used to identify the best predictor in regression models, etc. The package **StatMatch** offers two possibilities based on the computation of pairwise association measures (function pw.assoc) or on the analysis of uncertainty (this topic will be discussed in Section 4.1).

The guiding principle in the choice of the matching variables should be parsimony one: the larger is the subset of the matching variables the higher can be the undesired noise characterizing the final statistical outputs.

The function NND.hotdeck permits to search for the donors in a *constrained* setting, avoiding to select a donor more than once and by minimizing, at same time, the overall matching distance; in such case, the donors are identified by solving a traveling salesperson problem. The constrained matching returns an overall matching distance greater than the corresponding one in the unconstrained case, but it tends to better preserve the marginal distribution of the imputed variable.

***Random hot deck***

In the *random hot deck* the selection of the donors is performed completely at random, once the units have been divided in homogeneous groups according to the value of one or more common variables; for instance in surveys on people, units in both *A* and *B* can be grouped according to the gender, and consequently for a male in the recipient *A* it will be randomly selected a male in *B*.

The *random hot deck* method is implemented in the function RANDwNND.hotdeck; this function includes several variations of the traditional random hot deck. A very interesting feature consists in the possibility of handling non-fixed donation classes; in practice, a subset of potential donors is identified for each recipient unit and then, as usual, a donor is selected at random. This permits to use a continuous variable in creating the donation classes, without having to categorize it before the random selection. The Table 2 provides a summary of the alternative criteria made available in RANDwNND.hotdeck to identify the subset of the potential donors for each recipient unit.

**Criteria for identifying the subset of donors in random hot deck**

*Table 2*

| Donation classes | No. of donors in the subset | Value for argument `cut.don` | Value for argument $k$ |
|---|---|---|---|
| subset as a fraction $k$ of the closest donors | $n_D \times k$ | `cut.don="span"` | $0 < k \leq 1$ |
| Subset of the $k$ closest donors ($k$NN) | $k$ | `cut.don="exact"` | An integer $0 < k \leq n_D$ |
| Subset of the closest donors at a distance less or equal to $k$ | Variable | `cut.don="k.dist"` | A valued coherent with the chosen distance function |

$n_D$ denotes the overall number of available donors

As can be argued, the usage of a continuous $X$ variable in forming "moving" donation classes requires the computation of distances between recipient and potential donors. For this purpose it is possible to employ all the distance functions already mentioned for `NND.hotdeck`; in addition, by using the functions in the package RANN (Arya *et al.*, 2014) it is possible to perform a search of the $k$ Approximate Nearest Neighbors (ANN) of each recipient record. This search criterion is very efficient and fast even in the presence of a large number of units.

Recently it was added the chance of further reducing the set of potential donors by discarding the ones not satisfying a given inequality or equality constraint concerning the values of a variable that is observed in both the data sets. This permits

to introduce some auxiliary information in the SM procedure or to account for consistency rules when selecting donors. This criterion resulted very useful in the matching of the data of the Household Budget Survey (HBS) with those of the Survey on Income and Living Conditions (SILC) in order to investigate the relationship between household income and consumption expenditures in Italy (Donatiello *et al*, 2014). HBS provides detailed information on consumption and just a rough information concerning the income; this information cannot be directly compared with the one provided by the SILC survey, but was used as auxiliary information in the SM, so to avoid the CI assumption. In practice for each potential recipient in SILC, the subset of the potential donors in HBS is further reduced by considering just the households in the same or in adjacent income classes.

It is worth noting that the selection of donors can be carried out with probability proportional to weights associated to them (*weighted random hot deck*, cf. Andridge and Little, 2010). This procedure permits to account for the usage of weights (just the ones in the donor data set) in the SM application. The effects of such a choice are not fully explored in the SM context; simulation studies carried out by D'Orazio *et al.* (2012) ended with satisfactory results. Similar findings are achieved in Donatiello *et al.* (2014).

***Rank hot deck***

Singh et al. (1993) suggested a particular version of the nearest neighbor distance hot deck called *rank hot deck*; in practice the distance is computed on the percentage points of the empirical cumulative distribution function of the *X* variable being considered. The empirical cumulative distribution function is estimated by:

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} w_i I(x_i \leq x)}{\sum_{i=1}^{n} w_i} \qquad [2]$$

where $I(\ ) = 1$ if the condition within the parenthesis is satisfied and 0 otherwise; while $w_i$ is the weight assigned to the *i*th unit, e.g. the sampling weights (in absence of a weighting system $w_i = 1$ for all the units). Computing distances on the percentage points of the empirical cumulative distribution function avoids the problem of comparing directly the values of a variable affected by measurement errors where, however, errors do not affect the "position" of a unit in the whole distribution (D'Orazio *et al.*, 2006b). This method is implemented in the function rankNND.hotdeck which allows the usage of donation classes and the possibility of a constrained search of donors.

***Imputation of survey data with StatMatch***

The **StatMatch**'s functions implementing hot deck methods can be used for imputing the missing values in a survey without any particular adaptation, thus providing powerful tools also to survey experts not interested in performing statistical matching. The unique step required is the splitting of the survey data in two separate data sets: the recipient is the subset of the survey data presenting missing values on

the target variable, while the remaining units are put in the donor data set. All the enhancements implemented in the hot deck matching functions can be employed when imputing missing values, allowing researchers to use a wide set of distance functions and the possibility of selecting donors in a constrained setting, without having to write ad hoc code to solve the optimization problem. Moreover, the various enhancements introduced in the random hot deck may contribute to make the method attractive even in the context of imputation missing values where usually the nearest neighbor distance hot deck tend to be the preferred method.

## STATISTICAL MATCHING OF DATA FROM COMPLEX SAMPLE SURVEYS

Most of the SM methods are designed to integrate data from simple random samples. Often in National Statistical Institutes the data sources to integrate originate from complex sample surveys carried out on the same target population. In such cases, it is difficult to assume the i.i.d. to hold and therefore inferences should account for sampling design and the weights assigned to the units. However, sampling is just one of the sources of error in complex samples surveys; specialists in this area have to face *coverage errors*, unit and item *nonresponse errors* and *measurement errors*. Generally speaking the data set provided at the end of the whole survey process has fewer units than the planned ones because of coverage and unit nonresponse; it does not present item nonresponse because the missing values and values identified as wrong have been imputed; moreover the starting sampling weights have been modified to compensate for coverage and unit nonresponse.

SM methods that explicitly take into account the sampling design and the corresponding sampling weights are: Rubin's *file concatenation* (Rubin, 1986) and Renssen's approach based on *weights' calibrations* (Renssen, 1998).

The Rubin's approach consists in concatenating the initial data sets $S = A \cup B$ and re-calculating the sampling weights of the units (inverse of the probability of being included in the concatenated sample). This procedure poses several problems and, however, does not solve the problem of imputing the missing variables in the concatenated file; for this reasons the approach is seldom applied (see for instance Ballin *et al.*, 2008).

The Renssen's approach is more straightforward and easy to apply given that is based on weights' calibration, a widespread practice in sample surveys. The method is thought to provide an estimate of the frequencies in the contingency table $Y \times Z$ by exploiting all the information in the available data sources, including the units' weights. The procedure permits to estimate the final table under the CI assumption or by using an auxiliary data source *C*, in which *Y* and *Z* are jointly observed (a small additional survey, past survey, etc.). The additional data source *C*, containing all the variables (*X*,*Y*,*Z*) or just (*Y*,*Z*), can be exploited in two alternative ways: (a) by *incomplete two-way stratification*; and (b) *synthetic two-way stratification*. In practice, both the methods estimate $Y \times Z$ from *C* after some further calibration steps (for further details see Renssen, 1998).

A very appealing feature of the Renssen's procedure resides in its ability to maintain the coherence between the marginal distribution of the interest variables; in practice the estimated $Y \times Z$ table has marginal distributions of *Y* and *Z* coherent with the distributions estimated in the origin data sets, *A* and *B* respectively, after a first harmonization step. This preliminary harmonization step is aimed at aligning in *A* and *B* the joint/marginal distributions of the matching variables.

In **StatMatch** the harmonization of the matching variables can be performed by resorting to the function harmonize.x. On the contrary, the function comb.samples estimates the contingency table $Y \times Z$ with or without the additional data source *C*. When auxiliary information (contained in *C*) is not available, the table $Y \times Z$ is estimated under the CI assumption.

A recent option in comb.samples allows for imputation at micro level. In particular, for each unit in *A* it estimated the probability of assuming one the categories of *Z*, similarly, the probabilities of assuming a *Y* category are provided for each unit in *B*. These probabilities are obtained as a by-product of the whole procedure which is based on the fitting of *linear probability models*. These probabilities when used to estimate the marginal distribution of the imputed variable in the recipient data set (i.e. *Z* in *A*) return the same distribution estimated in the donor data set (*B* in our case). However, they have to be used with care, given that imputations are based on very basic models (linear probability models) which have well known drawbacks (estimated probabilities less than 0 or greater than 1, etc.).

## UNCERTAINTY IN STATISTICAL MATCHING

The absence of data where *Y* and *Z* are jointly observed is the major source of uncertainty concerning the matching results. When the objective of SM is macro, the uncertainty about the target parameters can be explored and it is possible to draw some conclusion about the target parameters. Let consider the basic matching framework being (*X*,*Y*,*Z*) categorical variables; the goal of SM consists in estimating the contingency table $Y \times Z$. This table cannot be estimated directly because *Y* and *Z* are not jointly observed; in the matching framework it is possible to estimate just their marginal distributions, respectively from *A* and from *B*. This information can be exploited to derive an interval of plausible values for probabilities in the table $Y \times Z$, by simply resorting to the *Fréchet classes*:

$$\max\left[0;\, P_{Y=j} + P_{Z=k} - 1\right] \leq P_{Y=j,Z=k} \leq \min\left[P_{Y=j}; P_{Z=k}\right], \qquad [3]$$

for $j = 1,2,\ldots,J$, $k = 1,2,\ldots,K$. This way of working does not end with unique estimate for each probability but with a set of equally plausible estimates given the starting data. The larger will be the intervals the higher will be the uncertainty in estimating $Y \times Z$ (D'Orazio *et al.*, 2006a).

In the presence of several common variables *X*, the width of the bounds, and consequently the uncertainty, tend to reduce when conditioning on the subset of the *X* variables that best predict *Y* and *Z*. In particular let consider the variable $X_D$

obtained as the cross-product of the chosen $X_M$ matching variables ( $X_M \subseteq X$ ), by conditioning on $X_D$ it comes out:

$$\max\left[0;\, P_{Y=j} + P_{Z=k} - 1\right] \;\le\; P_{j,k}^{(low)} \;\le\; P_{Y=j,Z=k} \;\le\; P_{j,k}^{(up)} \;\le\; \min\left[\, P_{Y=j}; P_{Z=k}\right] \quad [4]$$

where

$$P_{j,k}^{(low)} = \sum\nolimits_{i=1}^{I} \max\left[0;\, P_{Y=j,X_D=i} + P_{Z=k,X_D=i} - 1\right] \times P_{X_D=i} \quad [5a]$$

$$P_{j,k}^{(up)} = \sum\nolimits_{i=1}^{I} \min\left[\, P_{Y=j,X_D=i} \;;\; P_{Z=k,X_D=i}\right] \times P_{X_D=i} \quad [5b]$$

for $j = 1,2,\ldots,J$ , $k = 1,2,\ldots,K$ .

The probabilities involved in the computation of uncertainty bounds can be easily derived on the available data sources without integrating them at micro level; in particular $P_{Y=j|X_D=i}$ can be estimated from *A*, $P_{Z=k|X_D=i}$ can be estimated on *B*, while $P_{X_D=i}$ can be estimated on $A \cup B$ or simply on *A* or on *B*, assuming that both the data sources provide similar estimates as far as the marginal distribution of $X_D$ is concerned. If this is not the case, it would be preferable to harmonize the distribution of $X_D$ before estimating the probabilities.

The package **StatMatch** offers the possibility of estimating the Fréchet bounds for the cell probabilities in $Y \times Z$ by means of the function Frechet.bounds. cat. This function provides the uncertainty bounds conditioned or not on $X_D$. Moreover a summary measure of the uncertainty is derived by computing the average width of the uncertainty bounds:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} \left( \hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right) \quad [6]$$

The smaller is $\bar{d}$ the lower will be the uncertainty.

The exploration of uncertainty represents an approach to the SM that does not require maintaining a limiting assumption such as the CI. It provides a set of estimates that can be viewed as tool to support some decisions. For instance, the estimate of $P_{Y=j,Z=k}$ under the conditional independence assumption:

$$P_{Y=j,Z=k}^{(CIA)} = \sum\nolimits_{i=1}^{I} P_{Y=j|X_D=i} \times P_{Z=k|X_D=i} \times P_{X_D=i} \quad [7]$$

is always included in the uncertainty bounds (but it is not the midpoint of the interval) hence very short intervals will denote a situation with low uncertainty and where the CI assumption may be plausible. Moreover the uncertainty can be used as a tool for selecting the matching variables to be used in a "standard" SM application (in the sense that the problem is not tackled in terms of uncertainty)

***Choice of matching variables based on uncertainty***

As shown before in the case of categorical variables the width of the uncertainty bounds reduces by conditioning on $X$ variables that are highly associated with both $Y$ and $Z$. Therefore, in the presence of several common variables it seems logical to consider just the ones with the highest contribution to the reduction of the uncertainty. At this purpose the function Fbwidhts.by.x explores uncertainty in correspondence of each possible subset of the starting $X$ variables and summarizes it by means of the formula [6], then the researcher has to decide which subset of the $X$ seems more effective in reducing the uncertainty and therefore can be considered as the subset of the matching variables to be used in the matching application. This method for selecting the matching variables does not suffer of the drawbacks of other "traditional" methods; in particular it does not involve performing separate analysis on the available data sources and then combining in some manner the corresponding results. Unfortunately, the effort required by exploring uncertainty for all the possible combinations of the $X$ becomes non-negligible in the presence of several common variables. For this reason it is being developed a new automatic procedure (D'Orazio *et al.*, 2015) that attempts to limit the uncertainty exploration to the subset of the most relevant $X$ variables.

# CONCLUSIONS

The package **StatMatch** can be considered a mature package for statistical matching; the most widespread SM techniques are implemented. The directions for further improvements are strictly related to the main problems encountered in its application to data commonly available in National statistical offices. The areas which are likely of further developments are those related to the application of SM to data from complex sample surveys, to the exploitation of auxiliary information and to exploration of uncertainty. Improvement of methods and tools for performing SM with data arising from complex sample surveys is crucial in a national Statistical Institute, at this purpose the major focus is on the weights' calibration methods which seem more suitable to deal with data from household surveys (cf. Donatiello *et al*., 2015); in this context there is the need of further improving the available methods to better handle sources with mixed type variables and to improve the procedures to derive the final synthetic data sets.

## REFERENCES


1. Andridge, R.R., Little, R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse", International Statistical Review, 78, pp. 40-64.
2. Arya, S., Mount, D., Kemp, S. E., Jefferis, G. (2014). RANN: Fast Nearest Neighbour Search (wraps Arya and Mount's ANN library). R package version 2.4.1. http://CRAN.R-project.org/package=RANN
3. Ballin, M., Di Zio, M., D'Orazio, M., Scanu, M., Torelli, N. (2008), "File Concatenation of Survey Data: a Computer Intensive Approach to Sampling Weights Estimation", Rivista di Statistica Ufficiale, 2-3, pp. 5-12.

4. Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., Spaziani, M. (2014) "Statistical Matching of Income and Consumption Expenditures", International Journal of Economic Sciences, Vol. III, pp. 50-65.
5. Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., Spaziani, M. (2015) "The role of the auxiliary information in Statistical Matching Income and Consumption", paper presented at the New Techniques and Technologies for Statistics (NTTS) 2015 Conference, 10-12 March 2015, Brussels, Belgium.
6. D'Orazio, M. (2011a) "Statistical matching when dealing with data from complex survey sampling", in Eurostat, Report of WP1. State of the Art on Statistical Methodologies for Data Integration, ESSnet project on Data Integration, pp. 33-37. http://www.essnet-portal.eu/sites/default/files/131/FinalReport_WP1.pdf
7. D'Orazio, M. (2011b) "Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment" R package vignette. http://www.cros-portal.eu/sites/default/files//Statistical_Matching_with_StatMatch.pdf
8. D'Orazio, M. (2015) "StatMatch: Statistical Matching", R package version 1.2.3. http://CRAN.R-project.org/package=StatMatch
9. D'Orazio M., Di Zio M., Scanu M. (2006a) "Statistical matching for categorical data: Displaying uncertainty and using logical constraints". Journal of Official Statistics 22, pp. 137–157.
10. D'Orazio M., Di Zio M., Scanu M. (2006b) Statistical matching: Theory and Practice. Wiley, Chichester.
11. D'Orazio M., Di Zio M., Scanu M. (2012) "Statistical matching of data from complex sample surveys" Proceedings of the European Conference on Quality in Official Statistics - Q2012, 29 May - 1 June 2012, Athens, Greece
12. D'Orazio M., Di Zio M., Scanu M. (2015) "The use of uncertainty to choose the matching variables in statistical matching", paper presented at the New Techniques and Technologies for Statistics (NTTS) 2015 Conference, 10-12 March 2015, Brussels, Belgium.
13. Kovar J.G., MacMillan J., Whitridge P. (1988) "Overview and strategy for the Generalized Edit and Imputation System". Statistics Canada, Methodology Working Paper, No. BSMD 88-007 E/F.
14. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
15. Renssen, R.H. (1998) "Use of statistical matching techniques in calibration estimation", Survey Methodology, 24, pp. 171-183.
16. Rubin D.B. (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations". Journal of Business and Economic Statistics, 4, 87–94.
17. Särndal C.E., Swensson B., Wretman J. (1992) Model Assisted Survey Sampling. Springer-Verlag, New York.
18. 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, pp. 59-79.