This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Method: Statistical Matching Methods

Contents

General section

1. Summary

Statistical matching (SM) methods for microdata aim at integrating two or more data sources related to the same target population in order to derive a unique synthetic data set in which all the variables (coming from the different sources) are jointly available. The synthetic data set is the basis of further statistical analysis, e.g., microsimulations. The word synthetic refers to the fact that the records are obtained by integrating the available data sets rather than direct observation of all the variables. Usually the matching is based on the information (variables) common to the available data sources and, when available, on some auxiliary information (a data source containing all the interesting variables or an estimate of a correlation matrix, contingency table, etc.). When the additional information is not available and the matching is performed on the variables shared by the starting data sources, then the results will rely on the assumption of independence among variables not jointly observed given the shared ones.

The synthetic data set can be derived by applying a parametric or a nonparametric approach. They can be mixed too.

2. General description of the method

Statistical matching at micro level attempts to derive a synthetic data source by integrating the available data sources. In the traditional framework, there are two data sets $A=\{X,Y\}$ and $B=\{X,Z\}$, sharing a number variables *X* (common variables) while the variable *Y* is observed just in *A* and *Z* is available just in *B*. In practice the synthetic data source $S=\{X,Y,Z\}$ is derived by exploiting the shared information, i.e., the common variables *X* (usually a subset of them) and, when available, eventual auxiliary information concerning the relationship among *X*, *Y* and *Z* or just *Y* and *Z* which can be in terms of an additional data source in which all the variables are jointly observed or an estimate of a parameter of interest (correlation matrix, contingency table, etc.). It is worth noting that when the matching is solely based on the available common variables (*X*), then the results of the matching will rely on a strong assumption of conditional independence of *Y* and *Z* given *X*. Hence the entire analysis carried out on the synthetic data set will reflect such assumption (Chapter 2, D'Orazio et al., 2006)

From the practical viewpoint, the synthetic data set can be simply one of the origin data sources (*A* or *B*) in which the values of the missing variable are imputed using techniques developed for imputing missing values in a survey. Usually it is preferred to refer to the smaller data source (in terms of observations) which becomes the recipient; the other one, the larger data sets, plays the role of the donor. In some cases it may happen that the synthetic data set is the result of concatenating the original data sources ($S = A \cup B$), then two imputation steps are required, *Z* is imputes in *A* while *Y* is imputed in *B*. The file concatenation procedure is proposed by Rubin (1986) in order to deal with data arising from complex sample surveys carried out from the same target population. A similar procedure is suggested by Renssen (1998) whose approach, based on weights calibration, is essentially developed for macro purposes (estimation of two-way contingency table $Y \times Z$). A discussion about the methods for statistical matching data form complex sample surveys can be found in the Report of WP1 of the ESSnet on Data Integration (2011, pp. 43-49).

The methods that can be used to impute the values for the missing variable in the recipient data set (or the concatenated file) can be based on a parametric, nonparametric or mixed approach. For the sake of simplicity, it will be considered the case of two i.i.d. samples *A* and *B* and the conditional independence (CI) is assumed to hold.

2.1 Parametric approach

A model characterised by a finite number of parameters is explicitly considered; once its parameters are estimated it is possible to impute the values of the missing variables via conditional expectation (conditional mean matching) or by drawing values from the predicted distribution.

2.2 Nonparametric approach

Many applications of statistical matching are based on the usage on nonparametric methods which do not require specifying in advance a model. The most used nonparametric techniques in statistical matching derive from hot deck methods applied in sample surveys to fill in missing values. Usually the objective is that of creating the synthetic data set by imputing the missing variables in the recipient data set. Imputed values are those observed in a similar statistical unit observed in the donor data set. Random hot deck and nearest-neighbour hot deck are the most used techniques in statistical matching (cf. Section 2.4, D'Orazio et al., 2006). A discussion about the use of hot deck techniques is in D'Orazio et al. (2006) and Singh et al. (1993). Paass (1985) and Conti et al. (2006) enlighten that such methods may introduce a matching noise, i.e., a discrepancy among the joint probability density function of the variables of interest in the synthetic data set the ones in the target population.

2.3 Mixed methods

This class of techniques mixes parametric and nonparametric approach. More precisely, in a first step a parametric model is adopted and its parameters are estimated, then, in the second step, a completed synthetic data set is obtained by means of some hot deck procedures. This approach exploits the advantages of models, being more parsimonious as far as estimation is concerned, and, on the other hand, provides imputed values that are not artificial (i.e., predicted by the model with possibly a random term) but are really observed (taken from the donor records). Interesting papers in this context are those of Rubin (1986), Singh et al. (1993), Moriarity and Scheuren (2001, 2003).

3. Preparatory phase

Before integrating two data sources through statistical matching some practical steps are necessary (Chapter 3, ESSnet-ISAD, 2009):

i. identification of the common variables and harmonisation issues;

ii. choice of the matching variables

iii. Definition of a model (when using a parametric or mixed approach)

The harmonisation issue can be quite time consuming, because it may be necessary to harmonise the definition of units, the reference periods, the variables, the classifications etc. Sometimes harmonisation cannot be reached and if two variables available in both the data sources cannot be harmonised then they cannot be used as matching variables.

Once completed the harmonisation step, most of the matching methods listed in Section 3. require a crucial step for the choice of the matching variables X_M , i.e., the subset of the common variables ($X_M \subseteq X$) that should be used in the models or in computing distances among units. The commonly used approach to identify the set of matching variables consists in disregarding all those variables which are not statistically connected with *Y* or Z (Singh et al., 1988; Cohen, 1991). In this context it is possible to use methods commonly used to select the best subset of predictor when fitting regression models or nonparametric procedure based on the fitting of classification or regression trees. In the case of all categorical variables, D'Orazio (2011a) suggested a procedure which is based on the exploration of the uncertainty due to the statistical matching framework.

When it is necessary to specify a model it is worth noting that in the basic statistical matching framework since the variables (*X*,*Y*,*Z*) are not jointly observed on data, it is not possible to test the fit of the model to data. In this case, experts in the phenomena under investigation can provide guidance on the choice of the model. Another possibility consists in considering different alternative models where different results are evaluated (a kind of sensitivity analysis). It is worth noting that a mixed approach offers a certain level of protection against model misspecification if compared to a fully parametric one.

4. Examples – not tool specific

5. Examples – tool specific

Some statistical matching methods are implemented in a specific library, called "StatMatch" (D'Orazio, 2011b), made freely available for the R environment (R Development core team, 2014). As far as statistical matching at micro level is concerned the following functions are available:

(a) functions to perform nonparametric statistical matching at micro level by means of hot deck imputation (NND.hotdeck, RANDwNND.hotdeck, rankNND.hotdeck). The following examples taken from D'Orazio (2011a) show how to use the functions:

```
# example of usage of nearest-neighbour hotdeck
# with the R function NND.hotdeck

> group.v <- c("rb090","db040")
> X.mtc <- c("hsize","age")
> out.nnd <- NND.hotdeck(data.rec=rec.A, data.don=don.B,
+                       match.vars=X.mtc, don.class=group.v,
+                       dist.fun="Manhattan")

# to derive the systhetic data set
> fA.nnd.m <- create.fused(data.rec=rec.A, data.don=don.B,
+                          mtc.ids=out.nnd$mtc.ids,
+                          z.vars=c("netIncome","c.netI"))


# example of random hotdeck
# with the R function RANDwNND.hotdeck

> group.v <- c("db040","rb090")
> rnd.1 <- RANDwNND.hotdeck(data.rec=rec.A, data.don=don.B,
+                            match.vars=NULL, don.class=group.v)
> fA.rnd <- create.fused(data.rec=rec.A, data.don=don.B,
+                        mtc.ids=rnd.1$mtc.ids,
+                        z.vars=c("netIncome", "c.netI"))

> rnk.1 <- rankNND.hotdeck(data.rec=rec.A, data.don=don.B,
+                            var.rec="age", var.don="age")
> fA.rnk <- create.fused(data.rec=rec.A, data.don=don.B,
+                            mtc.ids=rnk.1$mtc.ids,
+                            z.vars=c("netIncome", "c.netI"),
+                            dup.x=TRUE, match.vars="age")
```

(b) a function to perform mixed SM at micro level for continuous variables (mixed.mtc); the following example is taken from D'Orazio (2011a):

```
> X.mtc <- c("Sepal.Length","Sepal.Width") # matching variables
# parameters estimated using ML
> mix.1 <- mixed.mtc(data.rec=iris.A, data.don=iris.B,
+                 match.vars=X.mtc,y.rec="Petal.Length",
+                 z.don="Petal.Width",method="ML", rho.yz=0,
+                 micro=TRUE, constr.alg="lpSolve")

> mix.1$filled.rec # provides A filled in with Z
```

6. Glossary

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

7. References

Cohen, M. L. (1991), Statistical matching and microsimulation models. In: Citro and Hanushek (eds.), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Vol II Technical papers*, Washington D.C.

Conti, P. L., Marella, D., and Scanu, M. (2006), Nonparametric evaluation of matching noise. *Proceedings of the IASC conference "Compstat 2006", Roma, 28 August – 1 September 2006*, Physica-Verlag/Springer, 453–460.

ESSnet on Data Integration (2011), *Report on WP1 State of the art on statistical methodologies for data integration*. http://www.cros-portal.eu/content/wp1-state-art

D'Orazio, M. (2011a), Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment. *R Package Vignette*. http://rm.mirror.garr.it/mirrors/CRAN/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf

D'Orazio, M. (2011b), StatMatch: Statistical Matching. R package version 1.0.3. http://CRAN.R-project.org/package=StatMatch

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical Matching, Theory and Practice*. Wiley, Chichester.

ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data (2009), *Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data*. http://cenex-isad.istat.it/

Little, R. J. A and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd Edition. Wiley, New York.

Moriarity, C. and Scheuren, F. (2001), Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* **17**, 407–422.

Moriarity, C. and Scheuren, F. (2003), A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* **21**, 65–73.

Paass, G. (1985), Statistical record linkage methodology: state of the art and future prospects. *Bullettin of the International Statistical institute, Proceedings of the 45th Session*, vol. LI, Book 2, Voorburg, The Netherlands.

R Development Core Team (2014), *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/

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

Rubin, D. B. (1986), Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business and Economic Statistics* **4**, 87–94.

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.

Specific section

8. Purpose of the method

Statistical matching (SM) techniques when applied at micro level aim at integrating the available data sources, related to the same target population, in order to derive a unique synthetic data set in which all the variables (coming from the different sources) are jointly available. The synthetic data set is the basis of further statistical analysis, e.g., microsimulations.

9. Recommended use of the method

1. Statistical matching techniques usually are applied to investigate the relationship between two variables, *Y* and *Z*, never jointly observed in the available data sources, by considering the available common information, usually *X* variables. When no auxiliary information is available the statistical matching is based on the conditional independence of *Y* and *Z* given *X*; unfortunately, this assumption cannot be tested on the available data. If the analyst does not consider it to be valid then the SM cannot be performed.

10. Possible disadvantages of the method

1.

11. Variants of the method

1. Parametric approach: conditional mean matching

   The conditional mean matching in the simple case of three continuous variables *X*, *Y*, and *Z* reduces to a regression imputation; the recipient data set *A* is filled in with the predicted values:

   $$\hat{z}_k^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_k, \qquad k = 1,2,\ldots,n_A$$

   The parameters of the model should be estimated in order to exploit all the available information in both the data sets. For instance, the simple estimation of the parameters obtained through their observed counterpart may lead to unacceptable results, like a non-positive semi-definite covariance matrix. A solution to this problem is to use the maximum likelihood estimation (cf. D’Orazio et al., 2006, pp. 16-19). A discussion of the problems concerning the combination of estimates obtained from the different data sets is in Moriarity and Scheuren (2001, 2003) and D’Orazio et al. (2006). Extension to this method to the multivariate case are provided in D’Orazio et al. (2006).

2. Parametric approach: stochastic regression imputation

   Regression imputation provides values lying on the regression and there is no variability around it. For this reason in the case of the previous example it would be better to refer to stochastic regression imputation, such that the imputed value is obtained as (cf. Little and Rubin, 2002):

   $$\tilde{z}_k^{(A)} = \hat{z}_k^{(A)} + e_k = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_k + e_k, \qquad k = 1,2,\ldots,n_A$$

being e_k a residual generated randomly from a normal distribution with zero mean and variance equal to the estimated residual variance $\hat{\sigma}_{Z|X}$. This is an example of the drawings based on the conditional predictive distributions. Extension to this method to the multivariate case are provided in D'Orazio et al. (2006).

3. Nonparametric approach: Random hot deck

   Random hot deck consists in randomly choosing a donor record in the donor file for each record in the recipient file. The random choice is often done within groups obtained by considering subsets of homogeneous units characterised by presenting the same values for one or more common variables *X* (usually categorical).

4. Nonparametric approach: nearest-neighbour hot deck

   Nearest-neighbour hot deck is widely used in the case of continuous variables. The donor unit is the closest to the given recipient units in terms of a distance measured by considering all or a subset of the common variables *X*. The distance can be measured in different ways (cf. Appendix C in D'Orazio et al., 2006). Sometimes the search of the donors is restricted to suitable subsets of the donor units, sharing the same characteristics of the recipient unit (as for random hot deck).

   The constrained nearest-neighbour hot deck represents an interesting variation of the nearest-neighbour hot deck. In this approach, each donor record can be chosen as donor only once: the subset of the donors to choose is the one obtained as a solution of the transportation problem whose objective is the minimisation of the overall matching distance (sum of the recipient-donor distances). This constraint helps in better preserving of the marginal distribution of the imputed variable in the synthetic data set.

   In general the methods based on distances pose the problem of deciding the subset of the common variables *X* to be used for computing it. Using all or too many common variables may affect negatively the matching results because variables with low predictive power on the target variable may influence negatively the distances.

5. Nonparametric approach: rank hot deck

   Singh et al. (1993) proposed the usage of the rank hot deck distance method; it searches for the closest donor for the given recipient record with distance computed on the percentage points of the empirical cumulative distribution function of the (continuous) common variable *X* being considered. Considering the percentage points of the empirical cumulative distribution provides values uniformly distributed in the interval [0,1]; moreover, this permits to compare observations when the values values of *X* cannot be directly compared because of measurement errors which however do not affect the "position" of a unit in the whole distribution.

6. Mixed approach: stochastic regression imputation followed by nearest-neighbour hot deck

   In case of continuous variables, the procedure resembles the *predictive mean matching* imputation methods; let *A* play the role of recipient then procedure follows these steps:

(step 1) Estimate (on *B*) the regression parameters of *Z* on *X*; then use the model to impute the predicted values of *Z* in *A* (it is preferable to add a residual error term to the predicted values);

(step 2) For each record in *A* impute the value of *Z* observed on the closest value in *B* according to a distance computed on the values of *Z* (predicted values of *Z* in *A* and truly observed values of *Z* in *B*).

Such a two steps procedure presents various advantages: it offers protection against model misspecification and also reduces the risk of bias in the marginal distribution of the imputed variable because the distances are computed on intermediate and truly observed values of the target variable, instead of a suitable subset of the common variables *X*. In fact when computing the distances by considering all the matching variables, variables with low predictive power on the target variable may influence negatively the distances. Various alternative similar mixed procedures are listed in D'Orazio et al. (2006, Section 2.5).

12. Input data

1. Ds-imput1: is the data set that contains data referred to the variables *X* and variable(s) *Y*, usually denoted as *A* in the statistical matching framework. This data set contains n_A records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on a given finite population *U*.
2. Ds-imput2: is the data set that contains data referred to the variables *X* and variable(s) *Z*, usually denoted as *B* in the statistical matching framework. This data set contains n_B records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on a given finite population *U*.
3. Ds-imput3: this is an optional data set that may be available in statistical matching as a source of auxiliary information. In such case it may contain all the necessary variables X, Y and Z or just Y and Z, the variables that are never jointly observed in the two basic input data sets (DS-input1 and ds-input2); usually this data set is denoted as *C* in the statistical matching framework and it contains n_C records (observations) usually representing a sample of i.i.d. observations or the results of a complex sample survey carried out on the same finite population *U* but in the past or on a smaller scale.

13. Logical preconditions

1. Missing values
    1. Usually the common variables are expected to be free of missing values and the same happens as far as the target variables are concerned. In some applications of nearest-neighbour hot deck it is possible to refer to distance functions that account for the missing values of the matching variables.
2. Erroneous values
    1.
3. Other quality related preconditions

    1.

4. Other types of preconditions

    1.

14. Tuning parameters

1.

15. Recommended use of the individual variants of the method

1.

16. Output data

1. Ds-output1: The output of the statistical matching at micro level is a synthetic data set in which all the interest variables *X*, *Y* and *Z* are available. The synthetic data set can be simply one of the origin data sources (ds-input1 or ds-input2) in which the values of the missing variables are imputed using methods listed before. Usually it is preferred to refer to the smaller data source (in terms of observations) which becomes the recipient; the other one, the larger data sets, plays the role of the donor. In some cases it may happen that the synthetic data set is the result of concatenating the origin data sources ($S = A \cup B$).

17. Properties of the output data

1.

18. Unit of input data suitable for the method

19. User interaction - not tool specific

1.

20. Logging indicators

1.

21. Quality indicators of the output data

1.

22. Actual use of the method

1.

Interconnections with other modules

23. Themes that refer explicitly to this module

1.

24. Related methods described in other modules

1.

25. Mathematical techniques used by the method described in this module

1.

26. GSBPM phases where the method described in this module is used

1. Phase 5 - Process

27. Tools that implement the method described in this module

1. R library *StatMatch* (D’Orazio, 2011b), made freely available for the R environment

28. Process step performed by the method

GSBPM Sub-process 5.1: Integrate data

Administrative section

29. Module code

Micro-Fusion-M-Statistical Matching Methods

30. Version history

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

31. Template version and print date

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