STATISTICAL MATCHING FOR DATA INTEGRATION WITH DIFFERENT DATA SOURCES

EVA XHAVA, INSTITUTE OF STATISTICS
exhava@instat.gov.al

Abstract

During the process of creating databases, which are frequently used by analysts and statisticians, several data files are combined by statistical matching techniques to enrich the host data file. This process requires the conditional independence assumption (CIA) which could lead to serious bias in the resulting joint relationships among variables. In this article, methods of statistical matching are considered. Results are based in real data from the file of Value Added Tax (VAT) and the file of Structured Trade Survey (STS), and for their implementation is used the language of programming R. Based on a case study, analysis of pre-requisites and the results of statistical matching methods are performed, and the gains from using auxiliary information are mentioned. Some conclusions about the methods of statistical matching performed are achieved. Specifically, it was confirmed that the CIA could be a serious limitation which could be overcome by the use of appropriate auxiliary information. Hot deck methods were found to be preferable to other methods performed in this case study. This article was motivated by the need to analyze together turnover data collected by different data sources, and in this case from the file of VAT and STS.

KEY WORDS:

Statistical matching; Auxiliary information; Conditional independence assumption

1. INTRODUCTION

Statistical matching is a model-based approach for providing joint information on variables and indicators collected through multiple sources[1] (surveys drawn from the same population). The potential benefits of this approach lie in the possibility to enhance the complementary use and analytical potential of existing data sources. Hence, statistical matching can be a tool to increase the efficiency of use given the current data collections. Often the aim of a matching exercise is to enlarge the information scope, but matching techniques have also been used for alignment of estimates observed in multiple surveys and for improving the precision of these estimates by integration with larger surveys. Two main approaches can be delineated in terms of outputs that can be obtained through matching:

a) The macro approach refers to the identification of any structure that describes relationships among the variables not jointly observed of the data sets, such as joint distributions, marginal distributions or correlation matrices (D'Orazio, 2006).

b) The micro approach refers to the creation of a complete micro-data file where data on all the variables is available for every unit. This is achieved by means of the generation of a new data set from two data sets that are based on an informative set of common variables between two 'synthetic micro records'. In practice, matching procedures can be regarded as an imputation problem of the target variables from a donor to a recipient survey. Y, Z are collected through two different samples drawn from the same population; X variables are collected in both samples and they are correlated with both Y and Z. The relation between these common variables with the specific variables observed only in one of the data sets - the donor data set will be explored and used to impute to the units of the other data set - the recipient data set - the variables not directly observed. Thus, a synthetic dataset is generated with complete information on X, Y and Z.

In particular, measures of association between Y and Z conditional on X cannot be estimated and they are usually assumed to be 0. This is the so called conditional independence assumption (CIA), a reference point for assessing the quality of estimates based on matching.
When this condition holds, matching algorithms will produce accurate estimates that reflect the true joint distribution of variables that were collected in multiple sources. It will give the same results as a perfect linkage procedure. Unfortunately, this assumption rarely holds in practice and it cannot be tested from the data sets. In case the conditional independence does not hold, and no additional information is available, the model will have identification problems and the artificial datasets produced may lead to incorrect inferences.

Another approach for tackling the conditional independence assumption is the use of auxiliary information. Auxiliary information usually comes in one of the following possible types:
a) Auxiliary parametric information, obtained from "hook"[2] variables.
b) A third data set (C) or an overlap of the two samples (A, B) that provides complete information on (X, Y, Z).

In a macro-matching parametric approach the auxiliary information, generally collected from hook variables, or through previous samples, archives or collection of data, can be particularly useful. Hook variables can contribute to significantly increasing the explanatory power of the common variables and therefore decrease the degree of uncertainty, and can eventually eliminate it completely in some cases. Auxiliary datasets can also be of use in the macro matching approach. The likelihood function can be split into two factors, and the data files A, B and C can be merged into one file. The final report of the ESSnet[3] on Data Integration identifies three main methodologies that focus on the use of auxiliary datasets with complete information:
• Singh et al (1993) proposes a two-step procedure for the use of auxiliary dataset in the context of "hot deck"[4] methods. First, a live[5] value of the variable Z from the data set C is imputed to each unit in data set A using one of the hot deck procedures. Secondly, for each record in A, a final live value from B will be imputed: the one corresponding to the nearest neighbour in B with a distance calculated on the previously determined intermediate value.
• Another methodology for the use of auxiliary information which takes into account complex sample designs is provided by Renssen (1998). Renssen identifies two approaches for providing estimates from the joint dataset, mainly focused on the adjustment of weights:
a) The 'calibration approach' that is obtained under

[1] Donald Rubin

[2] A "hook" variable is a variable in which can be saved a function or several functions which can be used in a special case from an existent program.
[3] European Statistical System
[4] The "hot deck" package contains all the necessary functions to do the hot deck imputations in a set of input data with missed observations using either the best cell method or the probabilistic method.
[5] The live value corresponds to the nearest neighbour

the incomplete two way stratification. This approach consists in calibrating the weights in the complete file (C) constraining them to reproduce in C the marginal distributions of Y and Z estimated from files to be matched.
b) A 'matching approach' where a more complex estimate of P (Y, Z) can be obtained under the synthetic two way stratification. Roughly speaking it consists in adjusting the estimates computed under the conditional independence assumption using residuals computed in C between predicted and observed values for Y and Z respectively.
• The third approach was proposed by Rubin (1986) and consists in appending the two data sources A and B. In the case of an overlap of samples, difficulties in estimating the concatenated weights can limit the applicability of this approach.

There is an important need to analyze together the data of turnover collected from different sources. There are two data files that have these data: Value Added Tax File (VAT) and Structured Trade Survey of Economic Enterprises file (STS). On the one hand, VAT file is an important source for the collection of turnover. On the other hand, STS collects a large range of variables relevant for economic analysis, and one of them number of employees.
This article aims to test the use of alternative model based techniques to integrate turnover information from VAT into STS file.

The objectives are:

Objective 1: Analysis of the coherence between turnover statistics based on currently collected STS turnover and VAT. This comparative overview of STS coherence with VAT shall provide important insights on the quality of the information collected in STS.
Objective 2: Assess the quality of turnover obtained through statistical matching in combination with variables collected in STS.

In the section 2 of the study are presented the main implementation steps of matching, highlighting the main results in relation to the two objectives. Meanwhile section 3 summarizes the main conclusions and recommendations for the application of statistical matching techniques in this article.

The question of the research is: Is there coherence between the statistics of turnover collected from VAT and STS?

In order to accomplish the objectives of the research, there should be tested the hypothesis on the absence or not of coherence between statistics of turnover collected from STS and VAT.

2. STATISTICAL MATCHING: METHODOLOGY AND RESULTS

The two data sources — VAT as donor and STS as recipient — share a set of common variables consistent in terms of definitions, classifications, marginal and joint distributions, and reference period. The two data sources have the same target population - employees.

2.1 METHODOLOGY

2.1.1 Comparison of distributions for common variables

On the basis of the selected target population (employees), the consistency of the marginal distributions of common variables is analyzed. The Hellinger distance metric (HD)[6] has been applied as a yardstick of similarity of distributions for the common variable used in the matching process. Below are presented the values of coefficients that compare two distributions of the common variable, NACE (the classification of activities according to the Nomenclature of Economic Activities, NACE Rev. 2).

```
$meas
tvd    overlap   Bhatt    Hell
 0         1       1       0
```

Dissimilarity index: The dissimilarity index is defined as the total variation distance (tvd) between the marginal distributions, and ranges from 0 (completely similar) to 1 (completely dissimilar). This index represents the fraction of records that are causing differences between the compared distributions. The smaller the dissimilarity index is, the more coherent the marginal distributions of the response variable in the donor and the integrated datasets are. Agresti suggests that as long as the dissimilarity rate is less than or equal to 6% (tvd ⩽ 0.06), the compared marginal distributions could be considered consistent.

Overlap: The overlap is the opposite of the dissimilarity index (sum of overlap and tvd is 1). Its value ranges from 0 (completely dissimilar) to 1 (completely similar). The higher the overlap is, the more coherent the compared marginal distributions are.

[6] In probability and statistics, Hellinger distance is used to measure the similarity between two probabilistic distributions.

[7] Hot deck is an imputation method that deals with data that are missed in which every missed value is substituted with an observed value in a survey from a similar entity.

To clarify, overlap and tvd are complimentary to each other and their sum is equal to 1. Analogously to Agresti's distributions' consistency suggestion (tvd ⩽ 0.06), it can be concluded that an overlap ⩾ 0.94 indicates that the compared distributions can be considered as consistent.
Hellinger's Distance: The Hellinger's distance is a dissimilarity index representing the distance between the two marginal distributions, which is non-negative, symmetric, and lies between 0 and √2. Hellinger's distance (Hd) is mathematically related to tvd by the following equation:

$$Hd^2 \leqslant tvd \leqslant Hd\sqrt{2}$$

Considering this equation and given that tvd ⩽ 0.06, one can derive that Hd ⩽ 0.042. In literature, when Hellinger's distance ⩽ 0.05 the two distributions are considered consistent.
Bhattacharyya Coefficient: The Bhattacharyya coefficient (Bhatt) is a measure of similarity between two distributions, and ranges from 0 to 1. This coefficient could be used to estimate the relative closeness of two distributions. The higher the value of the Bhatt coefficient is, the more similar the distributions are. The Bhatt coefficient can be mathematically related to the Hellinger's Distance through the following equation:

$$Hd = \sqrt{1 - bhatt}$$

Taking into account the limits of an acceptable Hellinger's distance (⩽ 0.05), the Bhatt coefficient would be acceptable if Bhatt ⩾ 0.9975.
In order to quantify the similarity between probability distributions of donor and recipient data, the Hellinger distance has been used, which lies between 0 and 1. Value 0 indicates a perfect similarity between two probabilistic distributions, whereas a value of 1 indicates a total discrepancy. Calibration techniques applied explained a perfect similarity for the common variable NACE, considering that Hellinger metric distance is equal to 0.

2.1.2 Analysis of the explanatory power for common variables

The choice of the matching variables is a crucial point in statistical matching. The conditional independence assumption is the reference point. The fulfilment of this condition guarantees that the joint distribution of matched variables Y and Z will be the same as the one obtained from a perfect linkage procedure.
The set of common variables is made by NACE code (Classification of Activities according to the Nomenclature of Economic Activities, NACE Rev. 2).

2.1.3 Matching methods

Often the goal is to obtain a complete synthetic micro data file through effective imputation of values to the unobserved variables.
The study is based on the data of the fourth quarter of the year 2016 and are tested several imputation methods like mixed methods for performing statistical matching and "hot deck "[7].
Mixed.mtc[8] function implements some mixed methods to perform statistical matching between two data sources that are showed below:

1. In the case of the estimation method under CIA (rho_YZ|X=0) where there are only parameter estimates (micro=FALSE), the estimated correlation matrix is as follows:

Table 1: The matrix of coefficients of correlations estimated under CIA

	NACE	STS120	Turnover
NACE	1.00000000		
STS120	0.04865750	1.000000000	
Turnover	0.04885995	0.002377403	1.00000000

In the table above STS120 refers to turnover in Structured Trade Survey of Economic Enterprises while Turnover refers to turnover in the file of VAT.

2. In the case of the estimation method with partial correlation coefficient (rho_YZ|X=0.5) where there are only parameter estimates (micro=FALSE), the estimated correlation matrix is as follows:

Table 2: The matrix of partial correlation coefficients where there are only parameters estimates

	NACE	STS120	Turnover
NACE	1.00000000		
STS120	0.04865750	1.000000000	
Turnover	0.04885995	0.501188700	1.00000000

3. In the case of the estimation method with partial correlation coefficient (rho_YZ|X=0.5) where there is an imputation step (micro=TRUE), the estimated correlation matrix is as follows:

Table 3: The matrix of partial correlation coefficients where there is an imputation step

	NACE	STS120	Turnover
NACE	1.00000000		
STS120	0.04865750	1.000000000	
Turnover	0.04885995	0.501188690	1.00000000

[8] This function implements some mixed methods to do statistical matching between two data sources.

4. In the case of Moriarity and Scheuren estimation method under CIA where there are only parameter estimates (micro=FALSE), the estimated correlation matrix is as follows:

Table 4: The matrix of partial correlation coefficients where there are only parameters estimates

	NACE	STS120	Turnover
NACE	1.00000000		
STS120	0.04869503	1.000000000	
Turnover	0.04889764	0.002377403	1.00000000

5. In the case of Moriarity and Scheuren estimation method with correlation coefficient equal with -0.15 (rho_YZ=-0.15), the estimated correlation matrix is as follows:

Table 5: The matrix of partial correlation coefficients in the case of Moriarity and Scheuren estimation method

	NACE	STS120	Turnover
NACE	1.00000000		
STS120	0.04869503	1.000000000	
Turnover	0.04889764	-0.150000000	1.00000000

NND.hotdeck function implements the distance hot deck method to match the records of two data sources that share the same variables. This function finds the closest donors computing Euclidean distance on NACE. It creates the synthetic data set filling STS with the turnover of VAT.

Because imputation approaches have usually limited ability to recreate individual level values, results are assessed in terms of preservation of important data distribution aspects and multivariable relationships (Rubin, 1996).

Therefore, to assess the robustness of different methods applied, it is compared the extent to which the observed distributions in the donor (VAT) are preserved in the recipient (STS) files. Hellinger distances are used again to measure the level of similarity of the joint distributions of turnover with key variables.

In a parametric framework, the assumption of conditional independence ensures that data are sufficient to estimate the parameters of the model.

2.2 RESULTS

The quality assessment in the context of matching needs a process approach. Each of the steps (the quality and the coherence of data sources, modelling techniques, matching/imputation algorithms) has a large impact on the quality of results.

When assessing results based on matched datasets, it should be considered the final aim of the analysis and their interpretation according to objectives of the study. Thus, results have to be interpreted in relation to the two-fold objective of the exercise.

Three main criteria were considered throughout the analysis:

1) The consistency of joint distributions (of turnover with matching variables) among
VAT observed, STS imputed and STS observed. The steps are as follow:
a) The comparison between VAT observed and STS observed helps for checking the coherence of common variables. In this particular case, it also helps to assess the quality of turnover information collected in STS with VAT as a benchmark taking into account objective 1.
b) The comparison between VAT observed and STS imputed serves as a quality criterion of matching referencing to objective 2.
c) The comparison between STS observed and STS imputed helps to compare how matching performs, in comparison with collected information in STS.

2) The consistency of different parameters such as totals, means, etc.

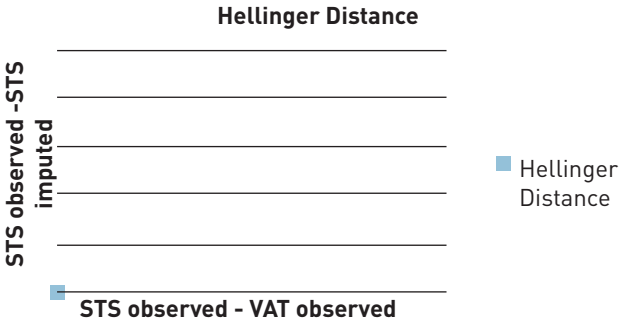
3) Test the CIA for specific target variables: turnover.

The objectives of the study are as follow:

OBJECTIVE 1 - Assess the quality of turnover information in STS with VAT as benchmark

OBJECTIVE 2 - Assess the quality of STS turnover statistics obtained through matching

Figure 1: Average similarity of joint distributions of turnover of STS



To assess the quality of results obtained through statistical matching we refer to two main criteria:

• Preservation of distributions and main parameters between the donor and the recipient.
• Capture the real joint distributions and correlations for variables not collected together.

After an analysis of the imputed turnover information, in general it is noted that the distributional parameters for the turnover variable as well as its joint distribution with matching variables are usually consistent between the donor (VAT observed) and the recipient (STS imputed). For instance, Figure 2 compares the cut-off points for turnover between VAT observed and STS observed for the fourth quarter of 2016. As it is seen, there are similar results both from matching and data collection.

The major limitation of statistical matching is its reliance on implicit assumptions. When imputed turnover need to be analyzed with additional variable collected solely in VAT, one essential condition for success is the existence of good explanatory variables that mediate the relation between these variables.

3. CONCLUSIONS AND RECOMMENDATIONS

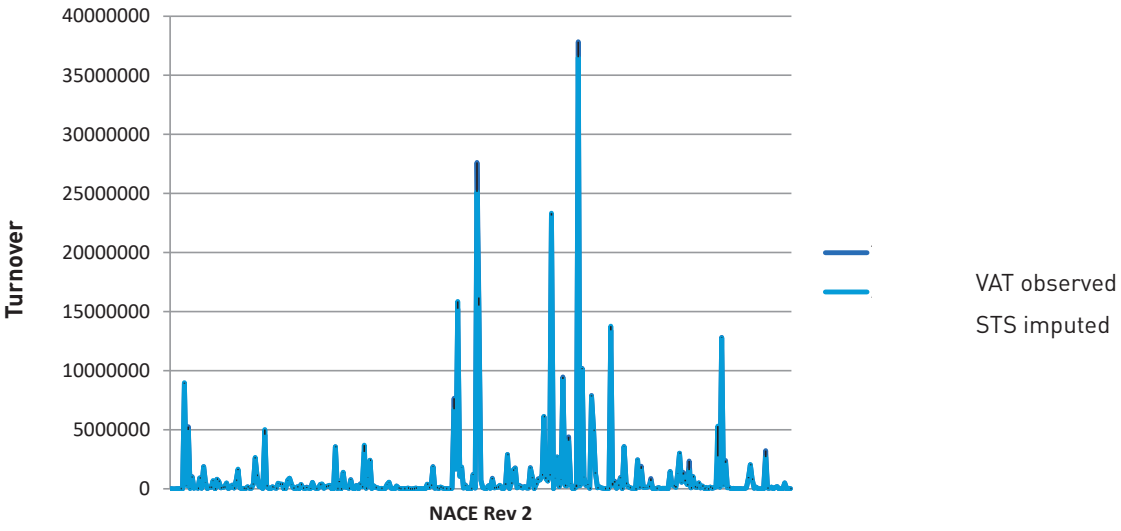
In this article, methods of statistical matching are considered. Results are based in real data from the file of VAT and STS, and for their implementation is used the language of programming R. Based in a case study are performed analysis of pre-requisites and of the results of methods of statistical matching, and are mentioned the profits from using auxiliary information. This article was based in two objectives and the conclusions are as follow:

OBJECTIVE 1: In basis of the analysis, the coherence of turnover of VAT and STS is good.

OBJECTIVE 2: An important factor for the joint analysis and matching of STS and VAT is a better coherence of variables. Differences and misalignment of distributions for the common variables used in the matching algorithm can cause discrepancies for turnover related estimates.

Specific matching methods proved to be more robust. However, results tend to be similar and in general estimates from matching are more sensitive to coherence pre-requisites and variables used in the model than the actual matching method employed.
Results show that, when pre-requisites of coherence are met, matching provides good results for marginal distributions and joint distributions that involve dimensions controlled in the model. However, when model assumptions hold, statistical matching can provide good inferences for specific estimates.
In the case studied, pre-requisites for matching are met. The two data sources - VAT as donor and STS as recipient – are consistent in terms of definitions, classifications, marginal distributions and reference period and the two data sources have the same target population.
Statistical matching should be used because it is a useful method in order to optimize the data sources. It allows using a small measure of sample using a priori stratified analysis with a smaller sample measures, compared with a sample not matched with a posteriori stratified analysis.
Matching avoids a stratified analysis with a lot of strata and indeed, in a not matched case study, while we do logistic regression, we may end up with empty strata.

Figure 2: Graphical illustration of VAT observed and STS imputed for the fourth quarter of 2016



BIBLIOGRAPHY

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2009) "The Use of Sample Weights in Hot Deck Imputation". Journal of Official Statistics, 25(1), 21–36.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

Andridge R.R., Little R.J.A. (2010) "A Review of Hot Deck Imputation for Survey Nonresponse". International Statistical Review, 78, 40–64.

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

D'Orazio M. (2010) "Statistical matching when dealing with data from complex survey sampling", in Report of WP1. State of the art on statistical methodologies for data integration, ESSnet project on Data Integration, 33–37, http://www.essnet-portal.eu/sites/default/files/131/ESSnetDI_WP1_v1.32.pdf

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, 137–157.

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, 137–157.

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, 137–157.

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

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

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

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2008) "The statistical matching workflow", in: Report of WP1: State of the art on statistical methodologies for integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 25–26. http://cenex-isad.istat.it/

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

D'Orazio M., Di Zio M., Scanu M. (2010) "Old and new approaches in statistical matching when samples are drawn with complex survey designs". Proceedings of the 45th "Riunione Scientifica della Societa' Italiana di Statistica", Padova 16–18 June 2010.

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.

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.

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.

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". Contributi Istat, 2005/10

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M. (2014), StatMatch: Statistical Matching (aka data fusion). R package version 1.2.2. http://CRAN.R-project.org/package=StatMatch.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), Statistical Matching: Theory and Practice. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

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

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

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

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

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

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

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

R¨assler S. (2002) Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. Springer Verlag, New York.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen R.H.(1998) "Use of statistical matching techniques in calibration estimation". Survey Methodology 24, 171–183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 171-183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 171-183.

Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation". Survey Methodology, No 24, pp. 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.

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

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

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, Journal of Business and Economic Statistics, 4, 87- 95.

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, Journal of Business and Economic Statistics, 4, 87- 95.

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, Journal of Business and Economic Statistics, 4, 87- 95.

Scanu M. (2008) "The practical aspects to be considered for statistical matching". in: Report of WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 34–35. http://cenex-isad.istat.it/

Scanu M. (2008) "The practical aspects to be considered for statistical matching". in: Report of WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 34–35. http://cenex-isad.istat.it/

Scanu M. (2008) "The practical aspects to be considered for statistical matching". in: Report of WP2: Recommendations on the use of methodologies for the integration of surveys and administrative data, "ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data", 34–35. http://cenex-isad.istat.it/

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, 59–79.

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, 59–79.

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, 59–79.

Singh, A.C., Mantel, H, Kinnack, M and Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption, Survey Methodology, 19, pp 59-79

Singh, A.C., Mantel, H, Kinnack, M and Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption, Survey Methodology, 19, pp 59-79

Singh, A.C., Mantel, H, Kinnack, M and Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption, Survey Methodology, 19, pp 59-79