ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

An R-Shiny Application for Statistical Matching

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Greco, F., Guastadisegni, L., Trivisano, C. (2025). An R-Shiny Application for Statistical Matching. Springer Cham [10.1007/978-3-031-64447-4_25].

Availability:
This version is available at: https://hdl.handle.net/11585/1004902 since: 2025-02-13

Published:
DOI: http://doi.org/10.1007/978-3-031-64447-4_25

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

18 November 2026

An R-Shiny Application for Statistical Matching

Fedele Greco, Lucia Guastadisegni, and Carlo Trivisano

Department of Statistical Sciences, University of Bologna.
fedele.greco@unibo.it, lucia.guastadisegni2@unibo.it,
carlo.trivisano@unibo.it

Abstract. Statistical matching involves integrating different data sources with different units and a set of common variables. Applying statistical matching methods presents challenges for practitioners, including data pre-processing and the application of different statistical matching methods. To address this, we develop an R-Shiny application to simplify the integration of different datasets for users and enhance efficiency in statistical matching tasks. In this paper, we delineate the various features of the application, encompassing tabs for importing new data or selecting from available surveys, choosing the matching variable, conducting data wrangling, employing diverse statistical matching methods, and using the matched dataset for regression analysis.

Keywords: Complex data survey, data integration, interactive tools, statistical matching

1 Introduction

Statistical matching encompasses a class of statistical methods designed to combine different datasets that share a set of common variables, some variables that are not jointly observed, across disjoint sets of units [2]. A notable characteristic of statistical matching methods is that they allow us to explore relationships between variables that are not observed simultaneously. Integration of two or more data sources can follow either a micro approach, where the objective is the construction of a complete synthetic file, or a macro approach, where the objective is the estimation of the joint distribution function of the variables of interest that have not been observed jointly [2]. Practitioners dealing with the integration of diverse data sources face numerous challenges, including recoding or recategorization of shared variables, substitution of some variables with new ones, conversion of measurement units, alignment of reference periods, and handling of missing or inconsistent data [2]. Moreover, applying various statistical matching techniques and generating a synthetic dataset can be intricate tasks. Therefore, the objective of this work is to develop an R-Shiny [4] application to aid practitioners in all stages of statistical matching, from importing raw data to creating matched synthetic datasets. The application will also facilitate regression analysis using the matched data. Additionally, it will provide R code

for generating matched datasets from raw data, ensuring the reproducibility of the analysis.

The paper is organized as follows. In Section 2, we offer a concise review of prominent statistical matching methods. Section 3 outlines the software description, highlighting its diverse features. Finally, in Section 4, we present and discuss some concluding remarks.

2 Statistical Framework and Methods

Let Y and Z be two random variables (r.v.s) and $\boldsymbol{X} = (X_1, \cdots, X_P)^\top$ be a vector of r.v.s of dimension P. Consider two datasets A and B referred to n_A and n_B statistical units, respectively. For each unit, independent and identically distributed (i.i.d.) observations generated from the joint distribution $f(\boldsymbol{x}, y, z)$ are available. The units in A have Z missing and the units in B have Y missing, that is:

$$(\boldsymbol{x}_a^A, y_a^A)^\top = (x_{a1}^A, \cdots, x_{aP}^A, y_a^A)^\top, \qquad (\boldsymbol{x}_b^B, z_b^B)^\top = (x_{b1}^B, \cdots, x_{bP}^B, z_b^B)^\top, \quad (1)$$

where $a = 1, \cdots, n_A$ and $b = 1, \cdots, n_B$. Statistical matching operates under the hypothesis that A and B are sampled from the same target population. We consider statistical matching methods based on the *Conditional Independence Assumption* (CIA) between Y and Z given $\boldsymbol{X}$, with density [2]

$$f(\boldsymbol{x}, y, z) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) \cdot f_{Z|\boldsymbol{X}}(z|\boldsymbol{x}) \cdot f_{\boldsymbol{X}}(\boldsymbol{x}). \quad (2)$$

The CIA cannot be tested but makes the model identifiable. Some of the methods that can be adopted under the CIA are described in the next subsections. As a general rule for all methods, A is the recipient file and B is the donor file. Usually, $n_B \geq n_A$. The final matched dataset has n_A units and includes the $\boldsymbol{x}$, y, and $\tilde{z}$ variable values, where $\tilde{z}$ has been imputed.

2.1 Parametric Methods

Parametric methods assume a parametric form for (2), that is

$$f(\boldsymbol{x}, y, z; \boldsymbol{\theta}) = f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}; \boldsymbol{\theta}_{y|\boldsymbol{x}}) \cdot f_{Z|\boldsymbol{X}}(z|\boldsymbol{x}; \boldsymbol{\theta}_{z|\boldsymbol{x}}) \cdot f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta}_{\boldsymbol{x}}), \quad (3)$$

where $\boldsymbol{\theta}_{y|\boldsymbol{x}} \in \Theta_{y|\boldsymbol{x}}$, $\boldsymbol{\theta}_{z|\boldsymbol{x}} \in \Theta_{z|\boldsymbol{x}}$, $\boldsymbol{\theta}_{\boldsymbol{x}} \in \Theta_{\boldsymbol{x}}$. Concerning the macro approach, thanks to the CIA, the model is identifiable, and maximum likelihood (ML) estimates of the parameters can be obtained by maximizing the following log-likelihood [9]:

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\theta}|\boldsymbol{A} \cup \boldsymbol{B}) &= \prod_{a=1}^{n_{\boldsymbol{A}}} f_{\boldsymbol{X}Y}(\boldsymbol{x}_a, y_a; \boldsymbol{\theta}) \prod_{b=1}^{n_{\boldsymbol{B}}} f_{\boldsymbol{X}Z}(\boldsymbol{x}_b, z_b; \boldsymbol{\theta}) = \\ &= \prod_{a=1}^{n_{\boldsymbol{A}}} f_{Y|\boldsymbol{X}}(y_a|\boldsymbol{x}_a; \boldsymbol{\theta}_{Y|\boldsymbol{X}}) \prod_{b=1}^{n_{\boldsymbol{B}}} f_{Z|\boldsymbol{X}}(z_b|\boldsymbol{x}_b; \boldsymbol{\theta}_{Z|\boldsymbol{X}}) \times \\ &\times \prod_{a=1}^{n_{\boldsymbol{A}}} f_{\boldsymbol{X}}(\boldsymbol{x}_a; \boldsymbol{\theta}_{\boldsymbol{X}}) \prod_{b=1}^{n_{\boldsymbol{B}}} f_{\boldsymbol{X}}(\boldsymbol{x}_b; \boldsymbol{\theta}_{\boldsymbol{X}}). \end{aligned} \quad (4)$$

Conditional Mean matching After estimating (4), a complete synthetic dataset can be reconstructed. If Z is continuous

$$\tilde{z}_a = E(Z|\boldsymbol{X} = \boldsymbol{x}_a) = \int_Z z f_{z|\boldsymbol{x}}(z|\boldsymbol{x}_a; \boldsymbol{\theta}_{z|\boldsymbol{x}}) dz \qquad a = 1, \cdots, n_a. \tag{5}$$

When Z is normally distributed, this is a classic regression imputation method [6], that is

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\boldsymbol{\beta}}_{Z\boldsymbol{X}} \boldsymbol{x}_a^A \qquad a = 1, \cdots, n_a. \tag{6}$$

Draws based on conditional predictive distributions A limit of conditional mean matching is that it lacks variability when imputing values for observations with identical conditioning variables. To address this limit, random values can be drawn from $f_{Z|\boldsymbol{X}}(z|\boldsymbol{x}_a; \hat{\boldsymbol{\theta}}_{z|\boldsymbol{X}})$ for every $a = 1, ..., n_A$. When Z is normal this process is known as stochastic regression imputation [6]:

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\boldsymbol{\beta}}_{Z\boldsymbol{X}} \boldsymbol{x}_a^A + \epsilon_a \qquad a = 1, \cdots, n_a, \tag{7}$$

where ϵ_a is randomly generated from a normal with 0 mean and estimated residual variance $\hat{\sigma}^2_{Z|\boldsymbol{X}}$.

2.2 Non-parametric methods

To avoid misleading results based on incorrect parametric assumptions, some non-parametric methods have been proposed for statistical matching, both at the micro and macro level. We consider only methods at the micro level, also known as hot-deck methods [12].

Random hot deck method This method randomly selects a donor record from the donor file for each record in the recipient file. Sometimes, this selection is done within specific subsets of units within the donor files using one or a few categorical variables shared between both files, such as demographic variables, forming what are known as donation classes [12].

Rank hot deck method In the rank hot deck method, a single ordinal matching variable X can be used for selecting donors to assign to the records in A [11]. Units in both files are ranked separately based on the values of X. If $n_B = kn_A$, where k is an integer, files are matched by pairing records with the same rank. In cases where the sample sizes of A and B differ, matching is conducted through the empirical cumulative distribution function of the distribution of X in the recipient file and donor file

$$\hat{F}_A^X(x) = \frac{1}{n_A} \sum_{a=1}^{n_A} I(x_a \leq x), \quad x \in X, \qquad \hat{F}_B^X(x) = \frac{1}{n_B} \sum_{b=1}^{n_B} I(x_b \leq x), \quad x \in X. \tag{8}$$

Subsequently, each $a = 1, \ldots, n_A$ is linked to the corresponding record b^* in B such that ([2])

$$\left|\hat{F}_A^X(x_a^A) - \hat{F}_B^X(x_{b^*}^B)\right| = \min_{1 \leq b \leq n_B} \left|\hat{F}_A^X(x_a^A) - \hat{F}_B^X(x_b^B)\right|.$$

Distance hot deck method In the unconstrained distance method, when there is a single continuous matching variable X, the donor for the a-th record in the recipient file A is chosen such that [8]

$$d_{ab*} = |x_a^A - x_{b*}^B| = min_{1 \leq b \leq n_B} |x_a^A - x_b^B|. \tag{9}$$

In the constrained distance methods, each record in B can be chosen as donor only once ($n_a \leq n_b$).

The hot deck distance-based techniques offer versatility by enabling integration with a variety of metrics. For continuous variables, this includes the Manhattan, Mahalanobis, and Exact distances, while for variables of different natures, the Gower index is utilized.

2.3 Mixed Methods

Mixed methods employ a combination of parametric and non-parametric techniques. The initial step involves estimating a parametric model for $(\mathbf{X}, Y, Z)$, followed by the derivation of a complete 'synthetic' dataset using a non-parametric approach. If Y and Z are continuous, the method comprises the following steps [10]:

1. The regression parameters of Z on X are estimated on file B, and the regression parameters of Y on X are estimated on file A.
2. Intermediate values $\tilde{z}_a$ are computed as in Equation (6) or (7) for the units in A. Similarly, values $\tilde{y}_b$ are also computed for the units in B.
3. For each $a = 1, \ldots, n_A$, impute the value z_b^* corresponding to the nearest neighbour b^* in B with respect to a distance between intermediate values determined at step 2, $d((y_a, \tilde{z}_a), (\tilde{y}_b, z_b))$.

Various versions of the aforementioned method have been proposed in the literature (see for example [10] and [12]), each implying modifications to either point 2 or 3.

3 Software description

The different features of the application can be accessed through a navigation sidebar that contains multiple tabs, each corresponding to a different phase of the statistical matching procedure. More in details, it is composed by the following tabs:

1. Summary of surveys and selection of the recipient and donor files: the first tab allows the user to select the surveys to be used for statistical matching. Some surveys are included in the application to enable the user to train with various statistical matching methods and explore the application's potential. Brief descriptions of each survey and variables are provided. Moreover, users have the opportunity to import new surveys into the application. The recipient file (A) and donor file (B) are selected.
2. Selection of the matching variables: once the A and B files have been selected in the first tab, the second tab allows the user to select the matching variables, i.e., the variables that are common to the two surveys. These variables, denoted as $\boldsymbol{\mathcal{I}}_A$ and $\boldsymbol{\mathcal{I}}_B$, are not yet ready to be used for statistical matching because they could exhibit inconsistencies between the two datasets.
3. Data wrangling and harmonization: in this tab, inconsistencies among the variables $\boldsymbol{\mathcal{I}}_A$ and $\boldsymbol{\mathcal{I}}_B$ can be resolved. In detail, categorical variables can be harmonized to have the same categories, continuous variables can be transformed into categorical ones (for example, age in years can be transformed into age classes), some variables can be transformed into new variables (for example, individual-level variables can be summarized into household-level variables), measurement errors can be adjusted, and so on. The harmonized variables are now denoted as $\boldsymbol{\mathcal{X}}_A$ and $\boldsymbol{\mathcal{X}}_B$.
4. Application of matching methods: in this tab, the application of statistical matching methods is enabled. Some of the statistical matching methods outlined in Section 2 are implemented using the StatMatch R package [3]. Within this tab, the variable Z is selected from dataset B, and the variable Y from dataset A. Through the different statistical matching methods presented in Section 2, the variable Z is imputed into dataset A. Several columns $\tilde{z}_a$ can be generated, each corresponding to a different matching method. This facilitates comparison among the various matching techniques.
5. Employment of matched data for estimating regression parameters: the last tab of the application allows the user to estimate regression models using the matched dataset. This enables the comparison of different regression parameters obtained from various matching methods [5].

The application allows downloading the R code needed to perform all the statistical analyses necessary to obtain matched data starting from raw data, thereby fostering reproducibility.

The application is currently under development, but it will be available by the time of the SIS Conference 2024.

4 Conclusion

In this work, we presented an R-Shiny application for statistical matching. The application assists the user in all phases of statistical matching procedures, from data importation to the application of regression analysis on matched data. Currently, the application is limited to statistical matching methods based on the CIA. As future work, more recent methods based on auxiliary information [7]

could be included in the application.

Acknowledgements

This study was funded by the European Union - NextGenerationEU, in the framework of the "GRINS -Growing Resilient, INclusive and Sustainable project" (PNRR - M4C2 - I1.3 - PE00000018 – CUP J33C22002910001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

1. Burkard, R. E., Derigs, U.: Assignment and matching problems: Solution methods with FORTRAN-programs. 184. Springer Science & Business Media (2013). https://doi.org/10.1007/978-3-642-51576-7
2. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical matching: Theory and practice. John Wiley & Sons (2006).
3. D'Orazio, M.: StatMatch: Statistical Matching or Data Fusion. R package version 1.4.1. (2022).
4. Chang, W, Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B.: shiny: Web Application Framework for R, R package version 1.8.0.9000 (2024). https://github.com/rstudio/shiny, https://shiny.posit.co/
5. Hirukawa, M., Prokhorov, A.: Consistent estimation of linear regression models using matched data. Journal of Econometrics. 203(2), 344–358 (2018). https://doi.org/10.1016/j.jeconom.2017.07.006
6. Little, R. J.A., Rubin, D. B.: Statistical analysis with missing data. 73. John Wiley & Sons (2019).
7. Moretti, A., Shlomo, N.: Improving Statistical Matching when Auxiliary Information is Available. Journal of Survey Statistics and Methodology. 11(3), 619–642 (2023). https://doi.org/10.1093/jssam/smac038
8. Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement. 1(3), 325–362 (1972).
9. Rubin, D. B.: Characterizing the estimation of parameters in incomplete-data problems. Journal of the American Statistical Association. 69(346), 467-474 (1974).https://doi.org/10.1080/01621459.1974.10482976
10. Rubin, D. B.: Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics. 4(1), 87–94 (1986). https://doi.org/10.1080/07350015.1986.10509497
11. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: On methods of statistical matching with and without auxiliary information. Technical Report (1990).
12. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. Survey Methodology. 19(1), 59-79 (1993).