Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics

Meinfelder, Florian; Schaller, Jannik

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Meinfelder, F., & Schaller, J. (2022). Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics. *Journal of Official Statistics*, *38*(2), 509-532. https://doi.org/10.2478/jos-2022-0024

sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 509–532, http://dx.doi.org/10.2478/JOS-2022-0024

Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics

Florian Meinfelder[1] *and Jannik Schaller*[2]

Data fusion describes the method of combining data from (at least) two initially independent data sources to allow for joint analysis of variables which are not jointly observed. The fundamental idea is to base inference on identifying assumptions, and on common variables which provide information that is jointly observed in all the data sources. A popular class of methods dealing with this particular missing-data problem in practice is based on covariate-based nearest neighbour matching, whereas more flexible semi- or even fully parametric approaches seem underrepresented in applied data fusion. In this article we compare two different approaches of nearest neighbour hot deck matching: One, Random Hot Deck, is a variant of the covariate-based matching methods which was proposed by Eurostat, and can be considered as a 'classical' statistical matching method, whereas the alternative approach is based on Predictive Mean Matching. We discuss results from a simulation study where we deviate from previous analyses of marginal distributions and consider joint distributions of fusion variables instead, and our findings suggest that Predictive Mean Matching tends to outperform Random Hot Deck.

Key words: Statistical matching; missing data; predictive mean matching; nearest neighbour Imputation; missing-by-design pattern.

1. Introduction

Data fusion, also known as statistical matching, is a perfect example of secondary data analysis. The objective of a data fusion is to jointly analyse variables from (at least) two different data sources which were not jointly observed, and each of the data sources originally served a different purpose.

The studies of the National Statistical Institutes (NSIs) are often committed to a particular objective such as measuring, for example, consumption expenditure of private households, in great detail. If the need arises to incorporate and combine information from several objectives, data fusion is a standard method to provide a microdata source, where these different types of information are artificially joined on an individual or household level.

[1] University of Bamberg, Statistics and Econometrics, Feldkirchenstraße 21, 96052 Bamberg, Germany. Email: florian.meinfelder@uni-bamberg.de
[2] Federal Statistical Office of Germany (Destatis), Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany. Email: jannik.schaller@destatis.de
Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of the Federal Statistical Office of Germany. The authors are grateful to Ralf Münnich and five anonymous reviewers for their valuable comments on this paper. We further thank Eurostat for the permission to use their R code and Pierre Lamarche to share his unpublished article, which is currently under review. The corresponding author would like to thank the DFG Research Unit 2559 MikroSim (https://mikrosim.uni-trier.de) for financial support.

In 2009, the Stiglitz-Sen-Fitoussi commission (Stiglitz et al. 2009) published a report on welfare and its components, which led to various approaches among NSIs within the European Union to measure the dimensions 'income', 'consumption' and 'wealth' (ICW) as proposed by the commission. Countries which do not measure all three dimensions within a single official statistics data source have been exploring data fusion methods in order to provide a corresponding data base (see e.g., Donatiello et al. 2014; Uçcar and Betti 2016; Albayrak and Masterson 2017; Dalla Chiara et al. 2019). The proposed data fusion methods are largely based on the research conducted by the European Statistical Office (Eurostat) (Lamarche 2017, 2018) and several NSIs (see e.g., D'Orazio et al. 2018) on statistically matching data from EU-SILC (European Union Statistics on Income and Living Conditions) with data from the HBS (Household Budget Survey). By matching EU-SILC and HBS, Eurostat and the NSIs pursue the goal to provide joint information about the household income details observed from EU-SILC and the household consumption expenditures observed from HBS. The original focus of their analyses had been on preserving marginal distributions and one of the preliminary findings is that Random Hot Deck (RHD), a classical nearest neighbour matching technique, performs very well in terms of preserving marginal distributions (Webber and Tonkin 2013; Serafino and Tonkin 2017; Lamarche 2018). However, the preservation of marginal distributions does not give any hint about whether the joint distributions of the variables not jointly observed, income and consumption expenditures in our case, is adequately reproduced in the matched data file.

Our research connects by extending the analysis objective to investigating associations between different variables, and we will explain in the following, why RHD yields good results for marginal distributions, but not necessarily for conditional or joint distributions. In addition, we aim to emphasize with our research the importance of considering not only strictly non-parametric data fusion methods, typically based on covariate-based nearest neighbour matching, which currently seems to be the default approach to data fusion in practice, although parts of the literature also discussed parametric variants (see e.g., Van der Putten et al. 2002; Donatiello et al. 2014; Lamarche 2018). Nearest neighbour methods have some appealing properties, but since data fusion is a particular missing-data problem, we recommend a more flexible approach to exploring various imputation methods in applied data fusion settings.

Therefore, we investigate the properties of an imputation method called Predictive Mean Matching (PMM) (Rubin 1986) which was extended by Little (1988) (who also coined the term *Predictive Mean Matching*) to multivariate data situations. PMM is a semi-parametric method which uses a parametric (typically linear) model to establish the basis for the subsequent matching. This approach is compared as a data fusion method to RHD proposed by Eurostat, and can be seen as a proponent for more parametric alternatives.

While both, RHD and PMM, are based on nearest neighbour hot deck matching, the underlying principles are very different, as RHD matches on the combined distances of the covariates, jointly observed in both studies, whereas PMM matches on the combined distances of model-based predictions of the variables which are observed in only one of the two studies. While these two methods do not exhaust the plethora of available nearest neighbour matching variants, they can be considered as archetypical for what we like to refer to as covariate-based matching and model-based matching. In order to discuss benefits and drawbacks of both data fusion archetypes, we compare the performance of Random Hot Deck and Predictive Mean Matching within a simulation study. As an

extension to the research conducted by Eurostat and the NSIs, we investigate primarily joint distributions and associations between different variables, especially between the variables not jointly observed.

To meet our objective of comparing the data fusion performance of RHD and PMM, we structure our article as follows: Section 2 contains a general overview of data fusion, followed by an in-depth description of the two aforementioned algorithms in Section 3. We investigate the properties of RHD and PMM as data fusion methods within a simulation study based on Scientific Use Files (SUFs) data from the EU-SILC study which we modify to mimic a data fusion situation. The setup of this stimulation study is described in Section 4 and we discuss the corresponding results in Section 5. We conclude the findings of our research in the final Section 6.

2. Methodological Aspects of Data Fusion Scenarios

In this section we introduce the basic notation used throughout this article, and we introduce perspectives on data fusion from statistical literature as well as the practitioners' perspective which often relies on the identification of 'statistical twins' or nearest neighbours.

2.1. Theoretical Background

Following the suggestion by Rubin (1986) to consider data fusion as file concatenation leads to the particular missing-by-design pattern (see e.g., Rässler 2002, chap. 4), and Figure 1 displays this schematic pattern if we stack two originally independent data sources A and B. The blank parts are missing and the corresponding variables were initially not part of the particular study, that is, variables Z are not observed in the first original study A (upper part of the stacked data set), and variables Y are not observed in the second original study B (lower part of the stacked data set). In this respect, we denote variables which are observed in both studies as X in the following, and we further denote variables relevant for the analysis which are only part of study A (but unobserved in study B) as Y and, analogously, variables required for the analysis which are only observed in study B (but unobserved in study A) as Z.

A typical data fusion analysis objective is based on variables Y and Z, and from the schematic overview it is apparent that we need identifying assumptions for the joint distribution of $f(Y, Z)$. In most imputation variants, either fully parametric or matching-based, an implicit *Conditional Independence Assumption* (CIA) is made, which was first pointed out

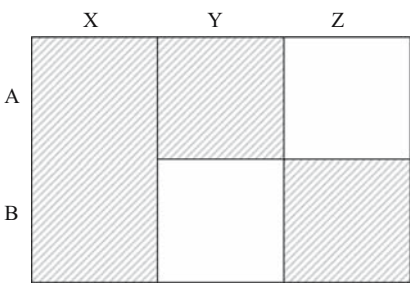

Fig. 1. Missing data pattern of a data fusion situation.

by Sims (1972) in a comment on a technical report (Okner 1972). It states that any association between Y and Z is a function of X, that is, $f(Y|X, Z) = f(Y|X)$ and, analogously, $f(Z|X,Y) = f(Z|X)$. This, for instance, yields a correlation of zero between Y and Z if conditioned on X. Furthermore, the CIA encases the *Missing at Random* (MAR) assumption and, therefore, also comprises ignorability (Koller-Meinfelder 2009). For details on different missing data mechanisms, see Little and Rubin (2020, chap. 1). We will, however, not consider violations of distributional assumptions for $f(Y, Z)$ as part of this research. Rodgers (1984) extensively discussed the shortcomings of the CIA within a comprehensive simulation study, but in recent years several publications have addressed this issue by proposing to introduce auxiliary information (see e.g., Singh et al. 1993; Zhang 2015; Fosdick et al. 2016). Imputation methods under non-ignorable missing data, however, have been discussed by Pfeffermann and Sikov (2011) and Little and Rubin (2020, chap. 15).

2.2. *Implementation in Practice*

Technically, we can apply any sophisticated method for handling missing data to this artificially created missing-by-design data situation, such as fully parametric multiple imputation or single imputation with variance correction, or Maximum Likelihood-based methods. The majority of empirical data fusions seem, however, to be based on variants of covariate-based nearest neighbour matching (see e.g., Koschnick 1995; D'Orazio et al. 2006a, sec 2.4), and we assume that there are at least two reasons for it:

(1) The synonymous term 'Statistical Matching' already insinuates matching-based methods to the practitioners as the most viable alternative (although this might not always be the case).
(2) The concept of identifying a 'statistical twin' makes data fusion appear as some kind of 'fuzzy' record linkage.

Although the missing-data pattern displayed in Figure 1 suggests that both missing parts could be imputed within the new stacked data set, it is far more common that only one of the original studies is used for data fusion analysis. Staying true to the matching concept, this study is labeled the *recipient* study, whereas we refer to the study that 'donates' data from its observations as the *donor* study (see e.g., Gabler 1997; Van der Putten et al. 2002). This applies to the content-based aspect of this article, where EU-SILC represents the recipient study that has to be extended by the missing consumption expenditures, while HBS donates the household consumption information and, therefore, serves as the donor study (see e.g., Serafino and Tonkin 2017). This implies that, referring to the missing-by-design pattern displayed in Figure 1, study A equals EU-SILC with the observed income variables Y and study B corresponds to the HBS with the observed consumption variables Z. The aim is to expand the EU-SILC data file by the household consumption expenditures, that is, imputing the missing Z information in study A in order to provide a joint analysis of the income (Y) and consumption (Z) variables originally not jointly observed.

2.3. *Overview of Traditional Data Fusion Algorithms*

Traditionally, as already pointed out, data fusions are conducted using some form of covariate-based nearest neighbour matching methods (see e.g., Rodgers 1984; Koschnick

1995). These algorithms match data on observations that are as close as possible with regard to their common X characteristics, that is, by imputing the missing Z values in the recipient data file by the observed Z values of the most similar donor observation according to the common X variables (see e.g., Van der Putten et al. 2002; Kiesl and Rässler 2005). Usually, either a certain distance metric is applied that accounts for the different scale levels of X, for example, the distance proposed by Gower (1971), or all X variables will be categorised such that (alleged exact) matches can be identified in both data files (D'Orazio et al. 2006a, sec. 2.4). In the latter case, only zero distances between the X variables are considered. In addition, however, there also exist fully parametric approaches. Such methods are based on regressions of Z on X within the donor file and subsequently estimate the missing Z values in the recipient file by means of the computed regression parameters (see e.g., D'Orazio et al. 2006a, sec. 2.2; Gilula et al. 2006).

While coviarate-based methods are non-parametric as they are not subject to distributional assumptions, PMM can be considered as a mixed (or semi-parametric) method between coviarate-based algorithms and fully parametric approaches. Note that Eurostat discusses different data fusion methods in previous working papers as well, and refers to semi-parametric algorithms as mixed methods between non-parametric and parametric approaches (Leulescu and Agafitei 2013; Webber and Tonkin 2013; Serafino and Tonkin 2017). However, their 'mixed methods' are slightly different to the multivariate variant of PMM proposed by Little (1988), as they consider ranks to identify suitable matches.

Nearest neighbour methods also spawn various other interesting research problems, such as balancing the usage of donors using constrained matching (see e.g., Rodgers 1984; Rubin 1986) or selecting donors from k Nearest Neighbours (see e.g., Chen and Shao 2000; Andridge and Little 2010; Beretta and Santaniello 2016). For the sake of simplicity, we do not explicitly address these issues in the following, although the setup of our simulations could be extended accordingly.

3. Implemented Data Fusion Algorithms

In this section we will provide some details on the two methods, RHD and PMM, we want to compare within the subsequent Monte Carlo study, before we focus on describing the aforementioned differences of both methods in detail, and how these differences might affect the analysis results of our MC study.

3.1. Random Hot Deck (RHD)

In general, RHD randomly assigns observations from the donor file to observations of the recipient file. The missing Z values for each recipient record are then imputed by the corresponding Z values of its assigned donor observation. However, the random allocation between recipient and donor observations is usually carried out within homogeneous subgroups, for example only within the same gender category. Thus, in this example, female (male) donor observations can only be assigned to female (male) recipient observations (D'Orazio et al. 2006a, sec. 2.4.1).

We also apply RHD within homogeneous subgroups analogously to Eurostat (2018) and Lamarche (2018). Note that Lamarche (2018) is currently under review and, therefore,

unpublished yet. A preliminary and freely accessible version of this article is Lamarche (2017). For any specific variable Z_r (with $r = 1, \ldots, p$) of $\mathbf{Z} = (Z_1, \ldots, Z_p)$ stemming from the donor data file, the detailed fusion algorithm can be described as follows: First, in order to identify relevant matching classes that serve as homogeneous subgroups, all common variables $\mathbf{X} = (X_1, \ldots, X_p)$ that have a metric scale level are categorised. For example, age is transferred to rough age groups or income to income quintiles. Thus, all $\mathbf{X}$ variables are at most ordinal scaled and only zero distances between the recipient and the donor observations are allowed. Subsequently, a stepwise selection based on OLS regression of Z_r on $\mathbf{X}$ in the donor file is implemented in order to select common $\mathbf{X}$ variables that have an acceptable explanatory power for Z_r. Along the stepwise-selected X variables, an auxiliary variable is created that concatenates the respective values of X for each observation in the recipient and the donor file. This results in a stratum characteristic for each donor and recipient record that form the homogeneous subgroups. The random assignment between the donor and the recipient observations is only conducted within the same stratum, that is, every donor record is only permitted to be matched with a recipient record that has exactly the same (categorical) characteristics with respect to the stepwise-selected X variables (Eurostat 2018; Lamarche 2018).

In order to ensure enough donor observations ($s_{l,don}$) compared to recipient records ($s_{l,rec}$) for each stratum level l, the following threshold is set:

$$\frac{s_{l,rec}}{s_{l,don}} \geq c \cdot \frac{s_{rec}}{s_{don}},$$

where the constant c is set as a rule of thumb to $c = 3$ (see Lamarche 2018, 13). In the case of equal sample sizes of the recipient and the donor data file, the threshold means that a maximum of three recipients can be assigned to one donor. As long as 90% of the sample do not exceed this threshold, the stepwise-selected X variables are retained and the Random Hot Deck within each subgroup is performed. Otherwise, the process will be reiterated, with the maximum subset of the $\mathbf{X}$ variables to be selected by the stepwise selection being reduced by 1 for each iteration step. The maximum subset of $\mathbf{X}$ variables is controlled by the `nvmax` argument within the `regsubsets()` function from the `leaps` package (Lumley and Miller 2020). If there are still recipients left who cannot be assigned to a donor, a second round of allocation is conducted with $c = 2$ and without a tolerance specification (Eurostat 2018; Lamarche 2018).

3.2. *Predictive Mean Matching (PMM)*

Predictive Mean Matching is not frequently discussed as a dedicated data fusion method, but it has become popular as an imputation method in general, and is the default method for metric-scale variables in the *R* package *mice* (Van Buuren 2021). The method was first introduced by Rubin (1986) and Little (1988) for the simultaneous imputation of continuous variables. The basic idea is that for each missing value its 'predictive mean' (Little 1988, 291) based on regression (e.g., OLS) is compared with the predictive means of all observed values, and the predictive mean among the observed values with minimum distance serves as donor record, and its actually observed values are imputed.

The PMM algorithm for any specific variable Z_r (with $r = 1,\ldots,p$) of $\mathbf{Z}$ is as follows: First, as with RHD, relevant $\mathbf{X}$ variables are selected using a stepwise selection based on OLS regression of Z_r on $\mathbf{X}$. In contrast to RHD, all $\mathbf{X}$ variables can remain on their original scale level, that is, metric variables are not categorised (Meinfelder and Schnapp 2015). By means of the regression equation, which includes the stepwise-selected X variables, the predictive mean is then calculated for each observation in the recipient and the donor file. In case $p > 1$, the search for corresponding donor observations is performed using the Mahalanobis distance function as proposed by Little (1988):

$$D_{i,j} = \left(\hat{z}_i - \hat{z}_j\right)^T S^{-1}_{Z_r|\mathbf{X}} \left(\hat{z}_i - \hat{z}_j\right)$$

with $i = 1,\ldots,n_{rec}$ and $j = 1,\ldots,n_{don}$, where $\hat{z}_i$ corresponds to the predictive mean of the i-th observation from the recipient file and $\hat{z}_j$ corresponds to the predictive mean of the j-th observation from the donor file. $S^{-1}_{Z_r|\mathbf{X}}$ denotes the $p \times p$-dimensional inverse variance-covariance matrix of the residuals from the regression of Z_r on the stepwise selection subset of $\mathbf{X}$, by which the distance is weighted.

3.3. Conceptual Differences of the two Algorithms

Traditional co-variate-based nearest neighbour methods like Random Hot Deck assign per default equal weights to all X variables, without taking into account the explanatory power or any other definition of relevance regarding the specific variables. Note that there exist weighted variants of hot deck imputation methods, where (sampling) weights can be assigned to specific observation units (see e.g., Andridge and Little 2009, 2010). However, we refer to the weighting of certain X variables used for imputation and we are not aware of any theory-driven weighting approach for RHD. Sometimes, as is the case with the considered implementation of the RHD algorithm, a regression-based stepwise selection of relevant X variables precedes the distance computation. While the X variables selected via the stepwise algorithm might be adequate predictors for the Y and Z variables to be matched, their explanatory power is unlikely to be equally high. Technically, any variable selection to identify 'suitable' X variables equates a weighting process, where weights for X variables are either 0 or 1. Although covariate-based methods like Random Hot Deck are widely used, the issue of unequal explanatory power of the common X variables results in an optimization problem for identifying the 'best' matches.

PMM, on the other hand, uses an underlying (generalized) linear model, where the contribution of any covariate of X is mapped to the explanatory contribution with respect to Z. The extension by Little (1988) offers also a solution for multivariate Z, as predictive mean matches are weighted with the Mahalanobis covariance matrix of the residuals from the regressions of Z_1 to Z_p on X. The rationale behind is that the distance between a recipient and a donor predictive mean of a particular variable Z_r is less penalized if the explanatory power of the regression model for Z_r is low (and, vice versa, the better the model fit, the more the distance is penalized), and we get a weighted sum of squares for the p distances between the predictive mean vectors.

Therefore, the main difference of PMM compared to RHD is in the distance processing: The possibility of unequal explanatory power among the X variables is ignored by RHD.

PMM takes potential unequal explanatory power of the selected X variables into account in a mathematically concise way.

This may indicate a more sophisticated distance processing in favour of PMM and, consequently, we expect PMM to result in a better fusion performance than the RHD method, especially with respect to preserving joint distributions. The upcoming simulation study is designed to provide a differentiated and detailed scrutiny of this underlying hypothesis.

3.4. Extension to Multiple Imputation

NSIs predominantly use descriptive analyses. While we therefore focus on single imputation in the simulation study, it is still possible to extend RHD and PMM to Multiple Imputation (MI) (Rubin 1978, 1987). MI is a popular method for dealing with missing data in general if the analysis objective is of an inferential nature. The basic idea is to replace any missing value several times. This is necessary, because the problem of any single imputation approach without variance correction is that no distinction is made in the analysis phase between observed and imputed values. MI uses a Bayesian framework, where missing values are replaced by $M > 1$ draws from a posterior predictive distribution of the missing data given the observed data. In a simplified data situation with complete variable(s) X and partially observed variable(s) Y this yields

$$p(Y_{mis}|X, Y_{obs}) = \int p\left(\theta|X, Y_{obs}\right) p(Y_{mis}|X, Y_{obs}, \theta) d\theta, \tag{1}$$

where Y_{mis} and Y_{obs} denote the missing and observed part of Y and θ denotes the parameters of the data generating model (Rubin 1987, 160). Technically, we have to consider the joint distribution of the data and the missingness (e.g., represented by an indicator variable R), but it can be shown that assuming *Missing at Random* and *Distinctness* we can ignore the model for the missingness (see e.g., Little and Rubin 2020).

One problem in MI is that direct random draws from Equation (1) are usually not feasible and we have to draw from the observed-data posterior distribution $p(\theta|X, Y_{obs})$ followed by drawing from the predictive distribution of the missing data $p(Y_{mis}|X, Y_{obs}, \theta)$ conditioned on the observed-data posterior instead. The additional variance component comes into play by creating $M > 1$ draws (and, therefore, M data sets) for each missing value which yields M different estimators $\hat{\theta}^{(m)}$ with $m = 1, \ldots, M$. Aside from the regular sampling variance of the estimator, which is calculated for all M data sets and averaged to get $W = \frac{1}{M}\sum_{m=1}^{M} \hat{V}\left(\hat{\theta}^{(m)}\right)$ we have the additional variance component of the M different estimators $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}^{(m)} - \hat{\theta}^{MI}\right)^2$ where $\hat{\theta}^{MI}$ as the so-called MI estimator is the average over all M different estimators. This variance reflects the uncertainty due to missingness. Rubin's rules combine the two variance components to the total MI variance $T = W + (1 + M^{-1})B$. Several publications discuss approaches, where semi- and non-parametric procedures approximate draws from the observed-data posterior and the conditional predictive distribution to emulate proper MI (see e.g., Burgette and Reiter 2010).

For RHD as a non-parametric method the Approximate Bayesian Bootstrap approach by Rubin and Schenker (1986), which was refined by Kim (2002) and Parzen et al. (2005), could be used to generate multiple imputations.

For the MI extension of PMM we can either use a parametric posterior step (Van Buuren 2018) or replace it by Bayesian Bootstrapping (see Koller-Meinfelder 2009), whereas the conditional predictive draw for the imputation is replaced by the predictive mean matching step, which is non-stochastic, but mimics random draws from a distribution quite well due to the deviations of the matched donors from the model. Zhou (2014) has also used Bayesian Bootstrapping to account for complex survey designs, which is why we think that incorporating information on varying inclusion probabilities could be conducted within the posterior step. Alternatively, multilevel models can be used to incorporate complex survey designs (Quartagno et al. 2020).

Note that for descriptive analysis a single imputation using $\hat{\theta}$ in combination with drawing from $p(Y_{mis}|X, Y_{obs}, \theta = \hat{\theta})$ is usually more efficient than additionally drawing from $p(\theta|Y_{obs}, X)$. If the imputation model is a linear model, and $\theta = [\beta, \sigma^2]$ this procedure is known as stochastic regression imputation (see e.g., Little and Rubin 2020). Analogously, we use the OLS estimators for PMM throughout our work, rather than posterior draws from the Bayesian linear model.

4. Simulation Design

Since the motivation for the present analysis is derived from the findings published by Eurostat (Webber and Tonkin 2013; Serafino and Tonkin 2017; Lamarche 2018), our aim for the data basis of the simulation study is to stay as close as possible to the relevant official statistics data sources.

With respect to the analysis objective of our comparison between RHD and PMM, we deviate from the focus of the previous studies, where emphasis was mainly put on preserving marginal distributions of the donor study within the fused data source. Instead, we concentrate on bivariate associations between common variables X and fused variables Z as well as on the primary objective of any data fusion, the joint distribution of the not jointly observed variables Y and Z. As stated above, we expect PMM to preserve the correlations between the variables Y and Z as well as between X and Z better than RHD, since correlations suffer more from non-exact matches than marginal distributions. Therefore, our simulation study focuses on evaluating the performance of both algorithms with respect to associations among different variables.

4.1. Database

We conduct a Monte Carlo (MC) study based on Scientific Use Files (SUFs) of EU-SILC from 2015, as Eurostat and the NSIs also focus on matching of the EU-SILC and HBS data files from 2015. In order to ensure a sufficiently large surrogate population, we combine EU-SILC data for Germany ($N_{DE} = 12{,}861$) and France ($N_{FR} = 11{,}384$). This leads to a total number of $N = 24{,}245$ observations from which we draw $k = 1{,}000$ random samples that we subsequently split into two data files which serve as substitutes for EU-SILC as recipient and HBS as donor file. For our simulation purposes, all data are based on EU-SILC due to the necessity of knowing the 'true' joint distribution of the simulated Y and Z variables and their correlations.

It would also be possible to draw the data from random distributions such as the normal distribution. However, it is important to ensure that the data-generating process is not

based on distribution families, such as multivariate normal distribution, as this would jeopardise a fair assessment of the respective data fusion algorithms. For example, Random Hot Deck and Predictive Mean Matching might benefit differently from assumed distributions. Since PMM is a semi-parametric approach and thus partially subject to distributional assumptions, generating data from a multivariate normal distribution might imply an advantage in favour of PMM compared to the non-parametric RHD algorithm. In this respect, a simulation based on empirical data seems more appropriate.

From the underlying data base, EU-SILC SUFs from 2015, we choose seven X variables as common variables. In order to stay as close as possible to the intended data fusion of Eurostat and the NSIs, the chosen variables represent those common variables Eurostat had selected for their analyses (Leulescu and Agafitei 2013; Lamarche 2018). Table 1 shows an overview of the $p = 7$ variables $X_1, \ldots, X_7$ used in the upcoming simulation study as well as the respective value range and measurement level. It can be seen that for RHD we have to categorise all variables, whereas for PMM the variables age (X_2) and income (X_7) keep their original metric scale level.

The variable activity status (X_1) contains information about the types of employment (self-employed or non-self-employed, pensioner, unemployed, etc.) (Eurostat 2016, 285). Details concerning the generation and recoding of this variable can be found in the online supplemental material. The *population density level* (X_3) indicates the population density of the current residential area. *Dwelling type* (Eurostat 2016, 173) reflects the type of accommodation (residential building, flat, etc.) (Eurostat 2016, 173). As both variables, population density level (X_3) and dwelling type (X_4), are empty for the German EU-SILC SUF (probably due to confidentiality reasons), we imputed the values using `mice` (Van Buuren 2021) via single imputation. The *tenure status* (X_5) combines information about the ownership status of the housing unit (sole owner, tenant, etc.) and on the (classified) rental costs incurred in case of a rental contract (Eurostat 2016, 174, 181). The binary variable *main source of income* (X_6) distinguishes between (1) income from self-employment or non-self-employment, property, ownership and assets and (2) income from pensions, social benefits and other transfers (Eurostat 2013, 20, 27–28; Eurostat 2016, 7, 313–316, 322–336). Its generation and recoding is documented in the online

Table 1. Overview of the Common X Variables.

Common X variables	Range / Measurement level	
	RHD	PMM
X_1: Activity status of RP[a]	1 to 5 / categorical	1 to 5 / categorical
X_2: Age of RP[a]	1 to 8 / categorical	acc. X_2 / metric
X_3: Population density level	1 to 3 / categorical	1 to 3 / categorical
X_4: Dwelling type[b]	1 to 4 / categorical	1 to 4 / categorical
X_5: Tenure status	1 to 5 / categorical	1 to 5 / categorical
X_6: Main source of income[c]	1 to 2 / categorical	1 to 2 / categorical
X_7: Income	1 to 5 / categorical	acc. X_7 / metric

[a] RP: 'Reference person' (interviewed person of the household);
[b] Actual range 1 to 5, category 5 is empty.
[c] Here, the missing values also form a category (coded as 9);
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

supplemental material. *Income* reflects the 'total disposable household income' (Eurostat 2016, 209) and, for RHD, is recoded into five income quintiles.

For the specific variables Y and Z, which represent the income variables of EU-SILC (Y) and the consumption variables of HBS (Z) that are actually not jointly observed, we select $p_{silc} = P_{hbs} = 2$ substitutes each, that is, $Y = (Y_1, Y_2)$ and $Z = (Z_1, Z_2)$, from the database. This underlines the possibility that the univariate data fusion (imputing only *one* HBS variable) can also be performed in a multivariate setting with more than one specific HBS characteristic. It becomes obvious that an exact coverage of the specific variables Y and Z is only possible for the income variables from EU-SILC, because the database itself consists of EU-SILC. However, since both, statistical and methodological conclusions, are of interest, it is more important to ensure the same measurement level for the respective income and expenditure substitutes and, therefore, it is essential to select *metric* variables.

For Y_1 we choose the variable 'total disposable household income before social transfers including old-age and survivor's benefits' (Eurostat 2016, 209) and for Y_2 the variable 'interest, dividends, profit from capital investments in unincorporated business' (Eurostat 2016, 214). The variables 'total household gross income' (Eurostat 2016, 207) and 'total disposable household income before social transfers other than old-age and survivor's benefits' (Eurostat 2016, 209) are selected for Z_1 and Z_2. Table 2 displays an overview of the specific Y and Z variables used for the simulation study and the corresponding measurement level.

Note that the variables X_7, Y_1, Z_1 and Z_2 are different household income variables, while Y_2 reflects capital gains. For information on the specific income concepts, see Eurostat (2016, 207–211, 214–215). The high proportion of income variables is due to the data situation that is based on EU-SILC only (where no consumption information is available – which is the reason for the intended data fusion of EU-SILC and HBS). However, in a real data fusion scenario, we do not know the joint distribution for Y and Z. Therefore, we base our simulation study on EU-SILC only and select different metric variables as proxies for income (Y) and consumption (Z) to evaluate the joint distribution of Y and Z. However, many metric variables in the EU-SILC SUFs have a high proportion of missing values, are completely blank or contain an excessive proportion of zeros. Hence, the selected Y and Z variables consist of those metric characteristics where information losses due to a high number of missing values or many zeros are as small as possible.

Table 2. Overview of the Specific Substitute Variables for EU-SILC (Y) and HBS (Z).

Specific EU-SILC substitute variables (Y)	Measurement level
Y_1: Total disposable household income before social transfers including old-age and survivor's benefits	metric
Y_2: Interest, dividends, profit from capital investments in unincorporated business	metric
Specific HBS substitute variables (Z)	**Measurement level**
Z_1: Total household gross income	metric
Z_2: Total disposable household income before social transfers other than old-age and survivor's benefits	metric

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

4.2. Monte Carlo Study

Our MC study is structured as follows: First, we draw $k = 1{,}000$ random samples without replacement (Jackknife) from the data specified above. Subsequently, for each random draw we generate the specific missing data pattern underlying data fusion scenarios (see Figure 1) and impute the missing Z_1 and Z_2 values in the $k = 1{,}000$ simulated data files via RHD on the one hand and PMM on the other hand.

More specifically, each random draw leads to a simulated data file that represents EU-SILC with the observed variables $X = (X_1,X_2,X_3,X_4, X_5,X_6,X_7)$ and $Y = (Y_1, Y_2)$ without information on the Z variables as well as to a simulated data file that represents HBS with the observed variables $X = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ and $Z = (Z_1, Z_2)$ that in turn contains no information about the Y variables. 'Stacking' both data sources results in the specific missing data pattern displayed in Figure 1. We impute the missing Z_1 and Z_2 values in the simulated EU-SILC data file using the two proposed data fusion algorithms, RHD and PMM. Thus, the imputed Z values in the matched data file reflect an artificial distribution $\tilde{Z} = (\tilde{Z}_1,\tilde{Z}_2)$. After the imputation step the correlations between Y and $\tilde{Z}$ as well as between the metric X variables (X_2: age and X_7: income) and $\tilde{Z}$ are then calculated and compared to the true correlations known from the surrogate population with $N = 24{,}245$ individuals as described in Subsection 4.1. As an additional diagnostic besides correlations, we provide specific conditional means of $\tilde{Z}$ given Y in the Appendix.

Note that in empirical data fusion situations we can typically only compare $f(\tilde{Z}|\mathrm{X})$ and $f(X, \tilde{Z})$ from the fused recipient study with $f(Z|X)$ and $f(X, Z)$, respectively, from the donor study, although this only makes sense if both data sources are random samples from the same population. Meinfelder (2013) proposes a scatter plot consisting of correlations $\rho_{X\tilde{Z}}$ from the fused data file and p_{xz} from the donor study, and the R^2 from a fitted linear regression could serve as a quality measure in this context. The artificially created distribution $f(Y, \tilde{Z})$, however, is subject to the CIA. Kiesl and Rässler (2006) show that the Frechet-Hoeffding bounds specify a theoretical lower and upper limit for the true marginal joint cdf of any pair of variables (Y_i, Z_j). Only the introduction of auxiliary information can help to narrow down the range for associations between Y and Z (see e.g., Fosdick et al. 2016).

As mentioned in Subsection 3.4 we apply single imputation to reflect those analyses which we presume NSIs to mostly carry out with the matched data base, resulting in point estimates for the correlations. This process, from sampling to imputation to the computation of correlations is performed with $k = 1{,}000$ simulation draws.

In order to get a rough understanding of how sensitive the performance of RHD and PMM could be with regard to the sample sizes of the recipient data file (n_{silc}) and the donor data file (n_{hbs}), we vary the sample size n. Of particular interest is the extent to which an excessive number of donors ($n_{silc} \lll n_{hbs}$) compared to an equal recipient and donor ratio ($n_{silc} = n_{hbs}$) has an effect on the performance of both data fusion algorithms. Therefore, the MC simulation is performed twice using different sample sizes n_1 and n_2. For both simulation scenarios we consider $n_{1_{silc}} = n_{2_{silc}} = 400$ observations for EU-SILC. In the first simulation scenario we also assign $n_{1_{hbs}} = 400$ units to the HBS. However, in the second scenario we choose a significantly higher number of donor observations, namely $n_{2_{hbs}} = 3,600$ for the HBS data. This leads to a sample size of $n_1 = 800$ with $n_{1_{silc}} = n_{1_{hbs}} = 400$ for the first scenario as well as to a sample size of $n_2 = 4,000$ with $n_{2_{silc}} = 400$ and $n_{2_{hbs}} = 3,600$ for the second scenario.

The MC simulation is conducted using `R` (R Core Team 2021), and we use packages `StatMatch` (D'Orazio 2020) and `BaBooN` (Meinfelder and Schnapp 2015) for RHD and PMM, respectively.

5. Results

Since we start out with complete samples, where parts of the data are removed to mimic a data fusion scenario, we know the true parameter values even for those parameters pertaining to $f(Y, Z)$, but the fusion algorithms implicitly rely on the CIA, and the theoretically correct values under this assumption are displayed as additional benchmarks in the results. Note, however, that while data fusion requires assumptions regarding the joint distribution of Y and Z, the identification problem and the natural uncertainty arising from it are not the primary focus of our work, but has been covered by many other authors (see e.g., Kamakura and Wedel 1997; D'Orazio et al. 2006b; Kiesl and Rässler 2006; Conti et al. 2012; Fosdick et al. 2016; Endres et al. 2019).

5.1. *Correlations Between* Y *And* $\tilde{Z}$

As stated in Subsection 3.3, we expect PMM to outperform RHD for bivariate associations. Hence, PMM should be able to reproduce the unobserved correlations between Y and Z more accurately than RHD. Table 3 displays the correlations between the specific variables $Y = (Y_1, Y_2)$ and $Z = (Z_1, Z_2)$, resulting from the artificial population consisting of $N = 24,245$ observations. These population correlations are used as true parameters for graphical diagnostics throughout this section and for Bias and MSE. The correlations between Y_1 and Z_1 as well as the correlations between Y_1 and Z_2 (0.87 and 0.85) are relatively high, whereas for Y_2 and Z_1 as well as for Y_2 and Z_2 we observe moderate correlations (0.44 and 0.48).

Figure 2 displays the MC distributions of the estimated correlations over all $k = 1,000$ MC simulation draws with equal number of recipients and donors ($n_{silc} = n_{hbs} = 400$). For high original correlations of 0.87 and 0.85, we find convincing evidence that PMM is able to reproduce the true parameter values more accurately than RHD. While RHD never covers the immediate area around the true correlations of 0.87 and 0.85 – the respective maximum for RHD amounts to 0.79 for $\widehat{\text{corr}}(Y_1, \tilde{Z}_1)$ and 0.71 for $\widehat{\text{corr}}(Y_1, \tilde{Z}_2)$ – the distributions of the estimated correlations resulting from PMM for $\widehat{\text{corr}}(Y_1, \tilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \tilde{Z}_2)$ are considerably close to the original parameter values. This also becomes clear by looking at the respective means: PMM produces mean correlations for $\widehat{\text{corr}}(Y_1, \tilde{Z}_1)$ of 0.83 and for $\widehat{\text{corr}}(Y_1, \tilde{Z}_2)$ of 0.78 and thus comes on average very close to the original values of 0.87 and 0.85, respectively. RHD generates mean correlations of 0.57 for $\widehat{\text{corr}}(Y_1, \tilde{Z}_1)$ and 0.54 for $\widehat{\text{corr}}(Y_1, \tilde{Z}_2)$ that deviate more strongly from the

Table 3. True Parameters for ρ_{YZ}

Corr (Y_1, Z_1)	Corr (Y_1, Z_2)	Corr (Y_2, Z_1)	Corr (Y_2, Z_2)
0.8678	0.8536	0.4361	0.4831

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

observed original parameter values. Furthermore, PMM reproduces the correlation between Y_1 and Z_1 more accurately than the respective relationship between Y_1 and Z_2. For moderate original correlations between Y and Z of 0.44 and 0.48 it can be seen that the superior performance in favour of PMM is slightly lower but still present. The MC distributions over all $k = 1{,}000$ simulation draws illustrated in Figure 2 show for $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_1)$ and $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_2)$ that the estimated PMM correlations cover the area around the true parameters more frequently while RHD tends to underestimate them. Consequently, the mean of the estimated correlations for $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_1)$ and $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_2)$ resulting from RHD are negatively biased with 0.20 and 0.21, respectively, while PMM produces almost unbiased estimators with MC mean correlations of 0.39 and 0.40.

The second scenario of the MC study contains an excessive number of donors in order to investigate mitigating effects on the results, if the proposed matching methods can choose from a larger donor pool, that is, the overall sample size $n_2 = 4{,}000$ consists of $n_{hbs} = 3{,}600$ donors versus $n_{silc} = 400$ recipients. In general, no substantial change can be observed for the RHD results as we can see in the respective MC distributions illustrated in Figure 3. The correlations resulting from PMM with n_2 indicate, compared to the PMM correlations with n_1 in the first scenario, in terms of $\widehat{\text{corr}}\,(Y_1, \tilde{Z}_1)$ and $\widehat{\text{corr}}\,(Y_1, \tilde{Z}_2)$ slightly improved results, since the bulk of MC distribution of the PMM correlations is even closer

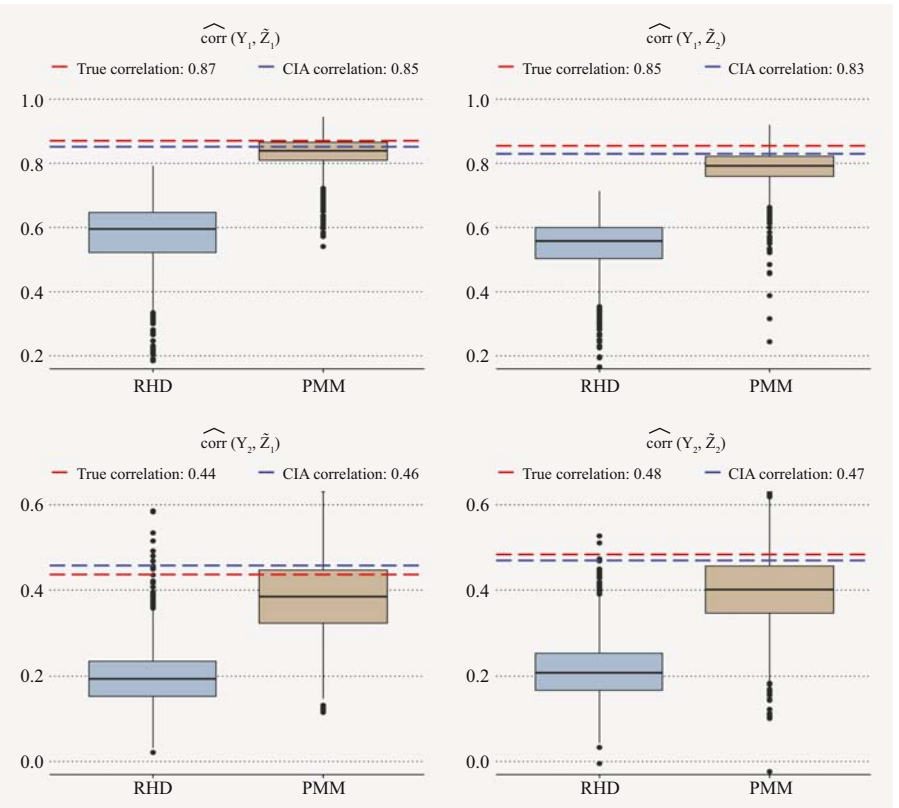

Fig. 2. Boxplots – MC distributions for $\hat{\rho}_{Y\tilde{Z}}$ with n_1.
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

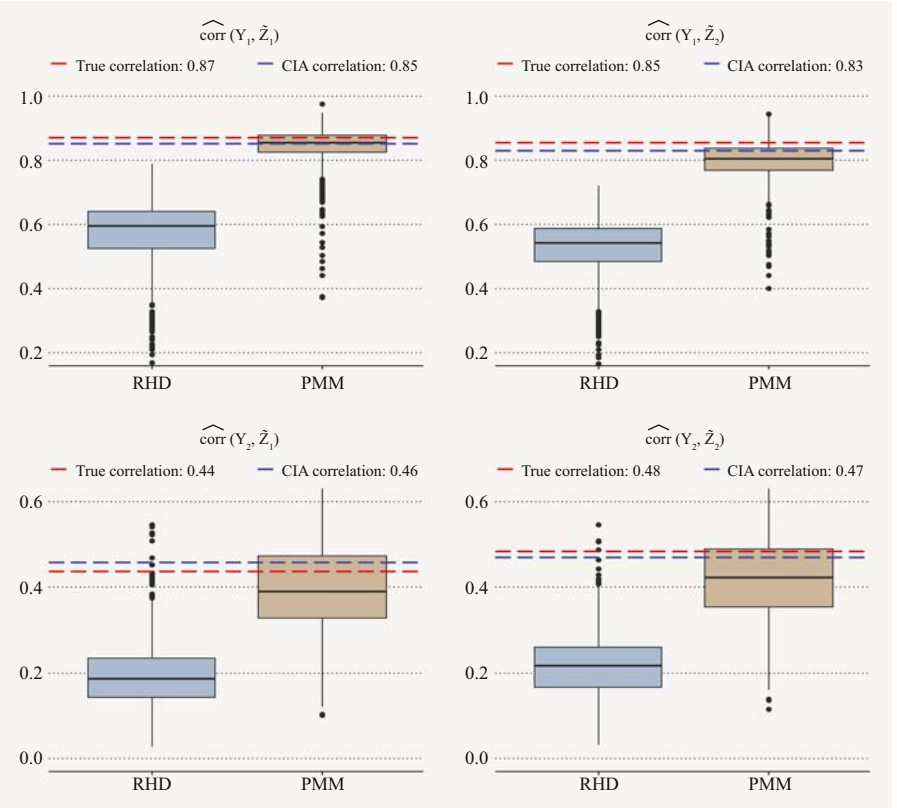

Fig. 3. Boxplots – MC distributions for $\hat{\rho}_{Y\tilde{Z}}$ with n_2.
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

to the true values of 0.87 and 0.85. Accordingly, the mean correlations computed with PMM under n_2 increase marginally to 0.84 for $\widehat{\text{corr}}\,(Y_1, \tilde{Z}_1)$ and 0.80 for $\widehat{\text{corr}}\,(Y_1, \tilde{Z}_2)$. In contrast, with respect to $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_1)$ and $\widehat{\text{corr}}\,(Y_2, \tilde{Z}_2)$ a slightly smaller number of the $k = 1,000$ PMM correlation estimates covers the area around the moderate true correlations of 0.44 and 0.48, while again the mean values marginally increase to 0.41 and 0.43 due to a higher outlier rate towards 1 that goes along with a somewhat higher variance. However, it should be noted that only small changes between n_1 and n_2 can be observed that should be treated with caution due to random fluctuations.

In addition to the true correlations of Y and Z, we marked the respective correlations under CIA in Figures 2 and 3, that is, the theoretical correlations of Y and Z, assuming independence if conditioned on X. This scenario underpins the relative advantage of PMM over RHD, as the PMM correlations are close to the correct values, whereas RHD fails to accurately reproduce the correlation structure of Y and Z, even though correlations under CIA are in this case close to the true correlations.

Consequently, the PMM correlation estimates have a smaller bias compared to the RHD estimates, as displayed in Table 4. While PMM yielded more outlier results within the MC simulation study, the MSE for PMM is still much lower than RHD's with respect to all four correlations of interest over both scenarios (see Table. 5).

Table 4. Bias of $\hat{\rho}_{Y\tilde{Z}}$

		$\widehat{\text{corr}}\ (Y_1, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (Y_1, \tilde{Z}_2)$	$\widehat{\text{corr}}\ (Y_2, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (Y_2, \tilde{Z}_2)$
n_1	RHD	0.2935	0.3139	0.2389	0.2707
	PMM	0.0386	0.0710	0.0450	0.0801
n_2	RHD	0.2950	0.3260	0.2435	0.2652
	PMM	0.0233	0.0554	0.0283	0.0520

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table 5. MSE of $\hat{\rho}_{Y\tilde{Z}}$

		$\widehat{\text{corr}}\ (Y_1, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (Y_1, \tilde{Z}_2)$	$\widehat{\text{corr}}\ (Y_2, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (Y_2, \tilde{Z}_2)$
n_1	RHD	0.0964	0.1070	0.0623	0.0785
	PMM	0.0044	0.0094	0.0123	0.0158
n_2	RHD	0.0963	0.1139	0.0647	0.0760
	PMM	0.0040	0.0073	0.0141	0.0153

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

5.2. *Correlations Between* X *and* $\tilde{Z}$

Apart from the reproduction of the joint information between the variables not jointly observed, the preservation of the distribution between the common variables X and the specific variables Z can be regarded as a minimum requirement, as it does not rely on any identifying assumptions. Table 6 shows the true correlations ρ_{XZ} resulting from the data base that represents our artificial population specified in Subsection 4.1. For X we consider the metric variables X_2 and X_7. Here, correlations between X_2 and Z_1 as well as between X_2 and Z_2 are relatively low with -0.11 and -0.02, respectively, while correlations for X_7 and Z_1 as well as for X_7 and Z_2 are rather high with 0.97 each.

As can be seen in the respective MC distributions displayed in Figure 4, both methods struggle with preserving the low correlations between X_2 and both specific variables. One possible explanation is that variable X_2 was not included in the backward-deletion selected matching model, as variable X_7 explains variables Z_1 and Z_2 almost perfectly, thus accidentally creating an uncongeniality issue (Meng 1994). For the scenario with excessive donors this phenomenon vanishes which can be seen in Figure 5, because the larger sample size increased the probability of X_2 to remain in the underlying model.

As expected, PMM does once again much better in preserving high correlations, as the respective correlation estimates for $\widehat{\text{corr}}\ (X_7, \tilde{Z}_1)$ and $\widehat{\text{corr}}\ (X_7, \tilde{Z}_2)$ come very close to the true parameter values and amount on average to 0.95 under n_1 and approximately to the real parameter value of 0.97 under n_2. RHD produces mean correlations of 0.64 (n_1) and 0.65 (n_2) for $\widehat{\text{corr}}\ (X_7, \tilde{Z}_1)$ as well as 0.65 (n_1) and 0.64 (n_2) for $\widehat{\text{corr}}\ (X_7, \tilde{Z}_2)$, respectively, that fall far behind the real observed parameter values. Investigating the bias and the MSE underlines these findings in favour of PMM, as displayed in Tables 7 and 8.

Table 6. True Parameters for ρ_{XZ}

corr (X_2, Z_1)	corr (X_2, Z_2)	corr (X_7, Z_1)	corr (X_7, Z_2)
−0.1081	−0.0208	0.9699	0.9737

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

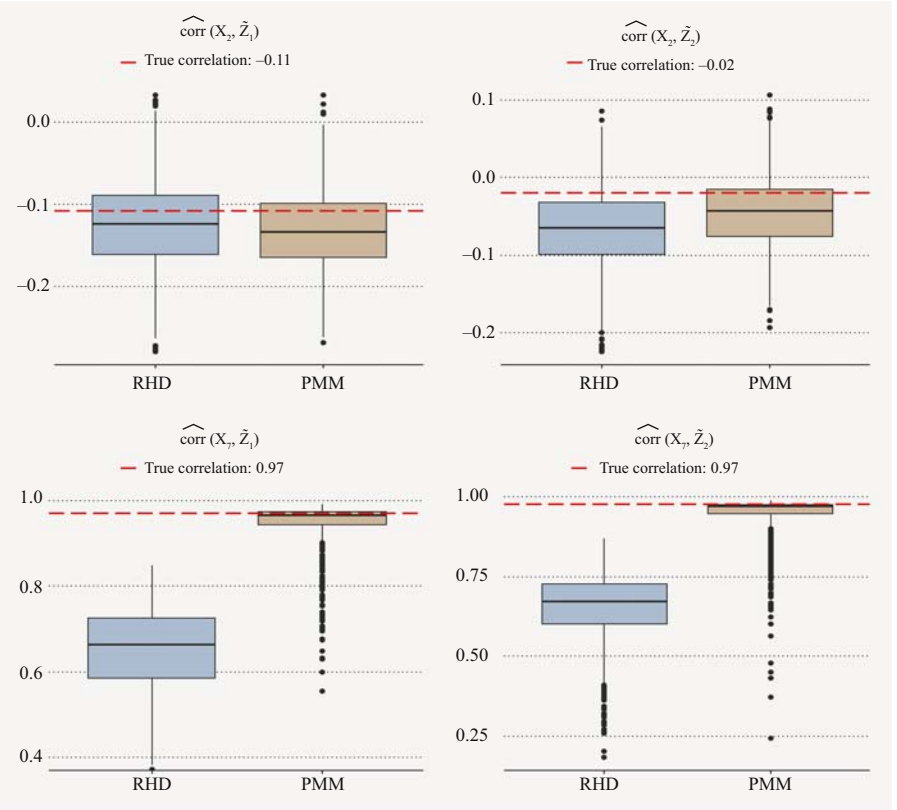

Fig. 4. Boxplots – MC distributions for $\hat{\rho}_{X\tilde{Z}}$ with n_1.
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

5.3. Discussion

Simulation studies can never claim general validity, but throughout the simulation study PMM emerged as superior method for the investigated quantities of interest, and there are several potential explanations for these findings.

The already mentioned benefit of PMM is the implicit weighting of the common X variables with respect to the specific Z variables. Since exact matches are rare in purely discrete settings and impossible in continuous settings, distances between recipients and potential donors play a crucial role. And some X variables usually turn out to be more relevant for the distance processing in order to find the 'nearest' donor observation and, therefore, we have to account for the unequal explanatory power of the common X variables with regard to the specific Z variables to be matched. Classical covariate-based methods do not provide a straightforward procedure to decide which variables should be included, but, using *all* potential common variables can lead to very inefficient matches.

For continuous X variables a perfect match is impossible which means that for the RHD method these variables have to be categorised which reduces information. An increasing number of variables and an increasing number of categories per variable means that the number of potential donors can become scarce or zero for some cell combinations and merging of categories is required for RHD which obviously additionally deprives the X

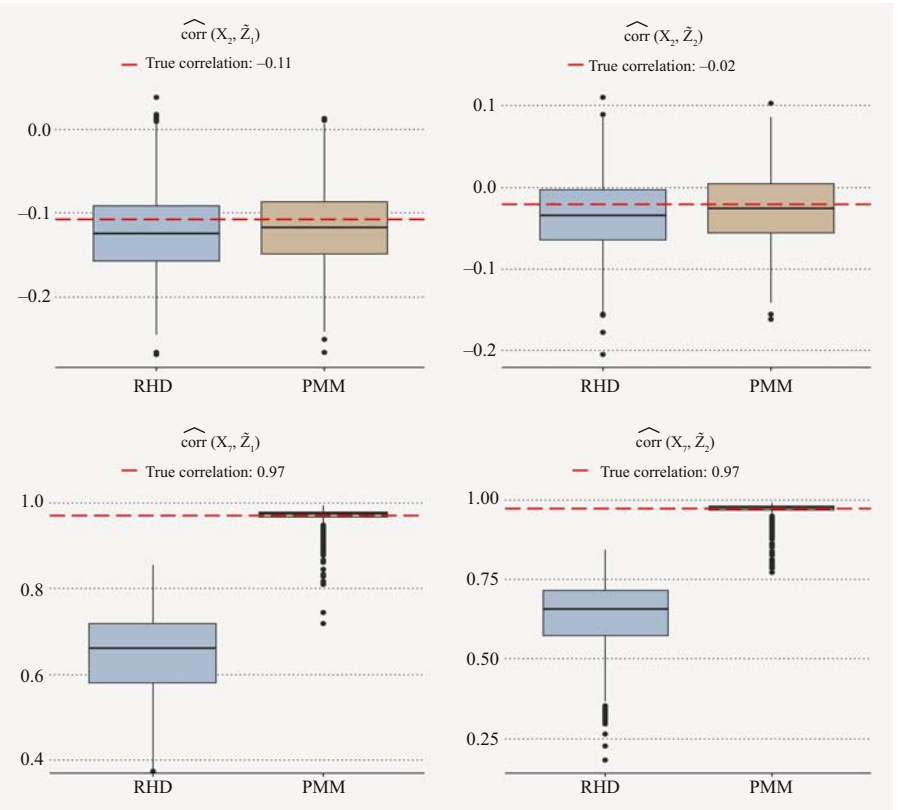

Fig. 5. Boxplots – MC distributions for $\hat{\rho}_{X\tilde{Z}}$ with n_2.
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table 7. Bias of $\hat{\rho}_{X\tilde{Z}}$

		$\widehat{\text{corr}}\ (X_2, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (X_2, \tilde{Z}_2)$	$\widehat{\text{corr}}\ (X_7, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (X_7, \tilde{Z}_2)$
n_1	RHD	0.0155	0.0449	0.3270	0.3264
	PMM	0.0240	0.0235	0.0236	0.0286
n_2	RHD	0.0145	0.0129	0.3299	0.3368
	PMM	0.0095	0.0059	0.0027	0.0043

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table 8. MSE of $\hat{\rho}_{Y\tilde{Z}}$

		$\widehat{\text{corr}}\ (X_2, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (X_2, \tilde{Z}_2)$	$\widehat{\text{corr}}\ (X_7, \tilde{Z}_1)$	$\widehat{\text{corr}}\ (X_7, \tilde{Z}_2)$
n_1	RHD	0.0032	0.0046	0.1203	0.1192
	PMM	0.0028	0.0027	0.0033	0.0055
n_2	RHD	0.0026	0.0022	0.1206	0.1255
	PMM	0.0022	0.0020	0.0005	0.0007

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

variables of some explanatory power. Generally speaking, when the probability for exact matches decreases due to a high number of common variables or a smaller donor pool, the way we are dealing with non-exact matches becomes more relevant.

PMM automatically handles this optimization problem which leads to overall better preservation of correlations between Y and Z, as matching on 'noise' is reduced. The usage of regression models means that PMM does not require any categorisation of X variables, but donor scarcity can still be an issue for PMM as well: The lower the number of potential donors, the higher, on average, the distances between recipients and matched donors (Andridge and Little 2010). This effect can also be observed for skewed variables (Kleinke 2017), where the long tail means that scarce donors are used more often which can also lead to biased results, although the potential for this problem is less severe in the large-scale studies in official statistics. Findings from Landerman et al. (1997) suggest that PMM remains an adequate imputation method even for income as a typically skewed variable that has been subject to our analyses as well. If, however, zero-distances are possible for all cell combinations, all nearest neighbour matching techniques are equivalent and, thus, RHD becomes a special case of PMM (see Little 1988, 291).

In order to investigate the aforementioned categorisation effects, we conducted additional simulations using a refined categorisation of both metric X variables for RHD with 14 age classes and 20 income classes. The results illustrated in the online supplemental material indicate an improvement of RHD, while PMM still outperforms RHD both for n_1 and n_2 significantly. Findings from additional simulations using Gower distances (without categorisation) suggest that PMM is still superior to covariate-based matching, although the gap becomes closer for the excessive donor scenario.

As stated in Section 3 in order to keep our research close to applied problems, we additionally included a variable selection scheme via backward-deletion for both algorithms to reduce the potentially high number of common variables to a more sensible subset of matching covariates. However, this does not invalidate the general perspective on the differences between PMM and RHD.

6. Conclusion

One objective of our research was to compare two types of data fusion methods, Random Hot Deck as a representative of 'classical' nearest neighbour hot deck methods, and Predictive Mean Matching as alternative. Covariate-based variants like RHD are widespread for data fusion in practice, whereas Predictive Mean Matching is a popular method for conditional univariate sequential regression imputation algorithms (Van Buuren and Groothuis-Oudshoorn 2011), but is far less common as a data fusion method. In general, PMM requires an underlying parametric model for predicting the (conditional) means of missing and observed cases, which distinguishes it from the purely covariate-based statistical matching methods. This can be perceived as a drawback, as the 'non-parametric' covariate-based methods do not require this step. However, we still make implicit assumptions for covariate-based algorithms about the association between the common and the specific variables by deciding to use a particular distance metric. Since our goal is the joint analysis of these variables, we assume a particular data generating process (including identifying assumptions), and we implicitly assume the imputation method to

be *congenial* to the analysis model, that is, associations among variables which are part of the analysis have to be accounted for by the imputation model as well (for details see Meng 1994; Xie and Meng 2017). To some extent this drawback therefore can be viewed as advantageous, as PMM requires us to think about the nature of the relationship between the common X variables and the specific Z variables to be fused.

We did not consider constrained matching (see e.g., Rodgers 1984; Rubin 1986) in this article which typically aims at a balanced usage of donors. We feel, however, that taking this approach to the extreme 'forces' the marginal distribution of Z from the donor study upon the recipient study, irrespective of different sample properties, indicated by deviating distributions in X. Under these circumstances it would be plausible that the fused distribution of $\tilde{Z}$ in the recipient study should be different to the corresponding distribution in the donor study.

Besides, the primary objective of any data fusion is the preservation of the joint distribution of Y and Z (Kiesl and Rässler 2005) which might clash with the objective of preserving the marginal distribution of Z at all costs if the studies are not random samples from the same population. In our simulation studies PMM outperformed the covariate-based RHD method with respect to preserving $f(Y, Z)$. A secondary objective of our research was to point out that covariate-based nearest neighbour matching should not automatically be considered as the default method for data fusion in practice. While Predictive Mean Matching can only be applied to metric-scale Z variables, we believe we have demonstrated that the method is a very useful addition to the toolbox of data fusion methods and, thus, should be taken into consideration for general application.

As stated previously, once data fusion is established as a particular missing data pattern with a particular analysis objective, we can leave the perspective of an artificially matched data set via 'statistical twins', and consider any sophisticated missing data method.

7. References

Albayrak, O., and T. Masterson. 2017. *Quality of statistical match of household budget survey and SILC for Turkey*, Levy Economics Institute, Working Paper (885). DOI: https://doi.org/10.2139/ssrn.2924849.

Andridge, R.R., and R.J.A. Little. 2009, "The Use of Sample Weights in Hot Deck Imputation". *Journal of Official Statistics* 25(1): 21–36. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-use-of-sample-weights-in-hot-deck-imputation.pdf (accessed March 2022).

Andridge, R.R., and R.J.A. Little. 2010. "A review of hot deck imputation for survey non-response". *International statistical review* 78(1): 40–64. DOI: https://doi.org/10.1111/j.1751-5823.2010.00103.x.

Beretta, L. and A. Santaniello. 2016. "Nearest neighbor imputation algorithms: a critical evaluation". *BMC medical informatics and decision making* 16(3): 74. DOI: https://doi.org/10.1186/s12911-016-0318-z.

Burgette, L.F., and J.P. Reiter. 2010 "Multiple imputation for missing data via sequential regression trees". *American Journal of Epidemiology* 172(9): 1070–1076. DOI: https://doi.org/10.1093/aje/kwq260.

Chen, J., and J. Shao. 2000. "Nearest Neighbor Imputation for Survey Data". *Journal of Official Statistics* 16(2): 113–131. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nearest-neighbor-imputation-for-survey-data.pdf.

Conti, P.L., D. Marella, and M. Scanu 2012. "Uncertainty analysis in statistical matching". *Journal of Official Statistics* 28(1): 69–88. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/uncertainty-analysis-in-statistical-matching.pdf.

Dalla Chiara, E., Menon, M., and F. Perali, F. 2019. "An Integrated Database to Measure Living Standards". *Journal of Official Statistics* 35(3): 531–576. DOI: https://doi.org/10.2478/JOS-2019-0023.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2014. "Statistical matching of income and consumption expenditures". *International Journal of Economic Sciences* 3(3): 50–65.

D'Orazio, M. 2020. *Statmatch: Statistical matching or data fusion: R-package*. Available at: https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf (accessed September 2021).

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. *Statistical matching: Theory and practice*, John Wiley & Sons.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints". *Journal of Official Statistics* 22(1): 137–157. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-matching-for-categorical-data-displaying-uncertainty-and-using-logical-constraints.pdf.

D'Orazio, M., Frattarola, D., A. Rizzi, A., M. Scanu, and M. Spaziani. 2018, *The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics*. Available at: https://www.istat.it/it/les//2018/11/Scanuoriginal-paper.pdf (accessed September 2021).

Endres, E., P. Fink, and T. Augustin. 2019. "Imprecise Imputation: A Nonparametric Micro Approach Reecting the Natural Uncertainty of Statistical Matching with Categorical Data". *Journal of Official Statistics* 35(3): 599–624. DOI: http://doi.org/10.2478/JOS-2019-0025.

EU-SILC SUF DE. 2015. *European union statistics on income and living conditions*. Scientific use file Germany. Available at: https://ec.europa.eu/eurostat/cros/EU-SILC-SUF_en.

EU-SILC SUF FR. 2015. *European union statistics*. Scientific use file France. Available at: https://ec.europa.eu/eurostat/cros/EU-SILC-SUF_en.

Eurostat. 2013. *European household income by groups of households*. Available at: https://ec.europa.eu/eurostat/documents/3888793/5858173/KS-RA-13-023-EN.PDF (accessed September 2021).

Eurostat. 2016. *Methodological Guidelines and Description of EU-SILC Target Variables: DocSILC065, 2015 Operation*. Available at: https://circabc.europa.eu/sd/a/afb4601b-4e5c-4f40-86bb-0c3d0d94aa12/DOCSILC065operation2015VERSION08-08-2016.pdf.

Eurostat. 2018. *R code to match EU-SILC and HBS*.

Fosdick, B.K., M. DeYoreo, and J.P. Reiter. 2016. "Categorical data fusion using auxiliary information". *The Annals of Applied Statistics* 10(4): 1907–1929. DOI: https://doi.org/10.1214/16-AOAS925.

Gabler, S. 1997. "Datenfusion". *ZUMA-Nachrichten* 21(40): 81–92.

Gilula, Z., R.E. McCulloch, and P.E. Rossi. 2006. "A direct approach to data fusion". *Journal of Marketing Research* 43(1): 73–83. DOI: https://doi.org/10.1509/jmkr.43.1.73.

Gower, J.C. 1971. "A general coefficient of similarity and some of its properties". *Biometrics* 27(4): 857–871. DOI: https://doi.org/10.2307/2528823.

Kamakura, W.A., and M. Wedel. 1997. "Statistical data fusion for cross-tabulation". *Journal of Marketing Research* 34(4): 485–498. DOI: https://doi.org/10.1177/002224379703400406.

Kiesl, H., and S. Rässler. 2005. "Techniken und Einsatzgebiete von Datenintegration und Datenfusion". In *Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung*, Tagungsberichte, Bonn: 17–32.

Kiesl, H., and S. Rässler. 2006. *How valid can data fusion be*? Available at: http://doku.iab.de/discussionpapers/2006/dp1506.pdf (accessed September 2021).

Kim, J.K. 2002. "A note on approximate bayesian bootstrap imputation". *Biometrika* 89(2): 470–477. DOI: https://doi.org/10.1093/biomet/89.2.470.

Kleinke, K. 2017. "Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching". *Journal of Educational and Behavioral Statistics* 42(4): 371–404. DOI: https://doi.org/10.3102/1076998616687084.

Koller-Meinfelder, F. 2009. *Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. PhD thesis, Bamberg. Available at: https://fis.uni-bamberg.de/bitstream/uniba/213/2/Dokument_1.pdf.

Koschnick, W.J. 1995. *Standard-Lexikon für Mediaplanung und Mediaforschung in Deutschland: Bd. 1.2, 2*., überarb. aufl. edn, Saur, München.

Lamarche, P. 2017. *Measuring Income, Consumption and Wealth jointly at the Micro-Level*. Eurostat. Available at: https://ec.europa.eu/eurostat/documents/7894008/8074103/income_methodological_note.pdf.

Lamarche, P. 2018. *Measuring Income, Consumption and Wealth jointly at the microlevel*. Eurostat.

Landerman, L.R., K.C. Land, and C.F. Pieper. 1997. "An empirical evaluation of the predictive mean matching method for imputing missing values". *Sociological Methods & Research* 26(1): 3–33. DOI: https://doi.org/10.1177/0049124197026001001.

Leulescu, A. and M. Agafitei. 2013, Statistical matching: A model based approach for data integration. Available at: https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF (accessed September 2021).

Little, R.J.A. 1988. "Missing-data adjustments in large surveys". *Journal of Business & Economic Statistics* 6(3): 287–296.

Little, R.J.A., and D.B. Rubin. 2020. *Statistical analysis with missing data*, third edition, John Wiley & Sons.

Lumley, T., and A. Miller. 2020. *leaps: Regression subset selection: R-package*. Available at: https://cran.r-project.org/web/packages/leaps/leaps.pdf (accessed September 2021).

Meinfelder, F. 2013. "Datenfusion: Theoretische implikationen und praktische umsetzung". In *Weiterentwicklung der amtlichen Haushaltsstatistiken*, edited by T. Riede, N. Ott, S. Bechthold, T. Schmidt, M. Eisele, B. Schimpl-Neimanns, F. Meinfelder, R. MŁunnich, J.P. Burgard and T. Zimmermann: 83–98.

Meinfelder, F., and T. Schnapp. 2015. *Baboon: Bayesian bootstrap predictive mean matching – multiple and single imputation for discrete data: R-package*. Available at: https://cran.r-project.org/web/packages/BaBooN/BaBooN.pdf (accessed September 2021).

Meng, X.-L. 1994. "Multiple-imputation inferences with uncongenial sources of input". *Statistical Science* 9(4): 538–558. DOI: https://doi.org/10.1214/ss/1177010269.

Okner, B. 1972. "Constructing a new data base from existing microdata sets: The 1966 merge file". In *Annals of Economic and Social Measurement*, 3(1): 325–362, National Bureau of Economic Research, Inc.

Parzen, M., Lipsitz, S.R., and G.M. Fitzmaurice. 2005. "A note on reducing the bias of the approximate bayesian bootstrap imputation variance estimator". *Biometrika* 92(4): 971–974. DOI: https://doi.org/10.1093/biomet/92.4.971.

Pfeffermann, D., and A. Sikov. 2011. "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information". *Journal of Official Statistics* 27(2): 181–209. Available at: https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/imputation-and-estimation-under-nonignorable-nonresponse-in-household-surveys-with-missing-covariate-information.pdf (accessed March 2022).

Quartagno, M., J.R. Carpenter, and H. Goldstein. 2020. "Multiple imputation with survey weights: a multilevel approach". *Journal of Survey Statistics and Methodology* 8(5): 965–989. DOI: https://doi.org/10.1093/jssam/smz036.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*, R. Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (accessed September 2021).

Rässler, S. 2002. "*Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*". Vol. 168 of Lecture notes in statistics, Springer, New York.

Rodgers, W.L. 1984. "An evaluation of statistical matching". *Journal of Business & Economic Statistics* 2: 91–102. DOI: https://doi.org/10.2307/1391358.

Rubin, D.B. 1978. "Multiple imputation in sample surveys – a phenomological bayesian approach to nonresponse". In *Proceedings of the Survey Research Method Section of the American Statistical Association*: 20–40. Available at: http://www.asasrms.org/GGT-SPU-f422b6f0b7825427-56279-110474-QWt4FYDtNN9fK3kX-LOD/Proceedings/papers/1978_004.pdf.

Rubin, D.B. 1986. "Statistical matching using file concatenation with adjusted weights and multiple imputations". *Journal of Business & Economic Statistics* 4(1): 87–94.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Rubin, D.B., and N. Schenker. 1986. "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse". *Journal of the American Statistical Association* 81(394): 366–374. DOI: https://doi.org/10.2307/1391390.

Serafino, P., and R. Tonkin. 2017. Statistical Matching of European Union Statistics on Income and Living Conditions (EU-SILC) and the Household Budget Survey, Eurostat. Available at:. https://ec.europa.eu/eurostat/documents/3888793/7882299/KS-TC-16-026-ENN.pdf (accessed September 2021).

Sims, C.A. 1972. "Comments (on Okner 1972)". *Annals of Economic and Social Measurement* 1: 343–345.

Singh, A.C., H.J. Mantel, M.D. Kinack, and G. Rowe. 1993. "Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption". *Survey Methodology* 19(1): 59–79.

Stiglitz, J., Sen, A., and J. Fitoussi. 2009. *Report of the Commission on the Measurement of Economic Performance and Social Progress* (CMEPSP). Avasilable at: https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf.

Uçar, B., and G. Betti. 2016. *Longitudinal statistical matching: transferring consumption expenditure from hbs to silc panel survey*, Technical report, Department of Economics, University of Siena. Available at: http://repec.deps.unisi.it/quaderni/739.pdf.

Van Buuren, S. 2018. *Flexible imputation of missing data*, CRC press.

Van Buuren, S. 2021. *Mice: Multivariate imputation by chained equations: R-package*. Available at: https://cran.r-project.org/web/packages/mice/mice.pdf (accessed September 2021).

Van Buuren, S. and K. Groothuis-Oudshoorn. 2011. "Mice: Multivariate imputation by chained equations in r". *Journal of Statistical Software* 45(3): 1–67.

Van der Putten, P., Kok, J.N., and A. Gupta. 2002. *Data fusion through statistical matching: Working paper 4342-02,* MIT Sloan School of Management. DOI: http://doi.org/10.2139/ssrn.297501.

Webber, D. and R. Tonkin. 2013. *Statistical Matching of EU-SILC and the Household Budget Survey to Compare Poverty Estimates Using Income, Expenditures and Material Deprivation*, Eurostat. Available: https://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF (accessed September 2021).

Xie, X. and X.-L. Meng. 2017. "Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial?". *Statistica Sinica*: l485–1545. DOI: https://doi.org/10.5705/ss.2014.067.

Zhang, L.-C. 2015. "On Proxy Variables and Categorical Data Fusion". *Journal of Official Statistics* 31(4): 783–807. DOI: http://doi.org/10.1515/JOS-2015-0045.

Zhou, H. 2014. *Accounting for Complex Sample Designs in Multiple Imputation Using the Finite Population Bayesian Bootstrap*, PhD thesis, Michigan. DOI: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.911.6156&rep=rep1&type=pdf.

Received November 2020
Revised May 2021
Accepted September 2021