*Journal of Survey Statistics and Methodology (2023) **11**, 619–642*

IMPROVING STATISTICAL MATCHING WHEN AUXILIARY INFORMATION IS AVAILABLE

ANGELO MORETTI*
NATALIE SHLOMO

There is growing interest within National Statistical Institutes in combining available datasets containing information on a large variety of social domains. Statistical matching approaches can be used to integrate data sources through a common set of variables where each dataset contains different units that belong to the same target population. However, a common problem is related to the assumption of conditional independence among variables observed in different data sources. In this context, an auxiliary dataset containing all the variables jointly can be used to improve the statistical matching by providing information on the correlation structure of variables observed across different datasets. We propose modifying the prediction models from the auxiliary dataset through a calibration step and show that we can improve the outcome of statistical matching in a variety of settings. We evaluate the proposed approach via simulation and an application based on the European Union Statistics for Income and Living Conditions and Living Costs and Food Survey for the United Kingdom.

KEYWORDS: Data fusion; Data integration; Distance hot deck; Model calibration; Predictive mean matching.

ANGELO MORETTI is an assistant professor in statistics at the Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands. NATALIE SHLOMO is professor in social statistics at the Social Statistics Department, University of Manchester, Manchester, United Kingdom

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 730998 (InGRID-2 Integrating Research Infrastructure for European Expertise on Inclusive Growth from Data to Policy).

*Address correspondence to Angelo Moretti, Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands; E-mail:a.moretti@uu.nl.

https://doi.org/10.1093/jssam/smac038 Advance access publication 13 February 2023

Statement of Significance

There is growing interest within National Statistical Institutes in combining available datasets containing information on a wide variety of social domains, that is, social exclusion, wellbeing, and poverty. Statistical matching approaches based on a common set of variables can be used when different units (e.g., households or persons) that belong to the same target population are contained in different data sources. However, a common problem is related to the conditional independence assumption that needs to be made to estimate the relationships among variables observed in different data sources. In this article, we use an additional auxiliary dataset to obtain the correlation structure of the relevant variables. We propose a calibration step in the prediction models for estimating the correlation matrices that improves the outcome of statistical matching, particularly when there are misspecification errors in the auxiliary dataset.

1. INTRODUCTION

National Statistical Institutes (NSIs) collect statistical information on a large variety of societal aspects, such as social exclusion, income, and living conditions. However, data on these topics are typically collected by different surveys where there are no common statistical units across surveys. If multiple data sources containing *different* statistical units (e.g., households, persons, and businesses) are available but belong to the same universe, one can combine the datasets using statistical matching approaches (D'Orazio et al. 2006; de Waal 2015). Statistical matching is an umbrella term describing the idea of fusing two or more datasets containing different statistical units.

Assuming the statistical matching involves two data sources, one dataset is defined as the recipient and the other one as the donor. The recipient file, denoted by A, contains a variable Y, which is not present in the donor, denoted by B, while variable Z is contained only in the donor file B. The goal of statistical matching is to use information contained in the set of common variables (also known as background variables) X, to match records from the donor to the recipient file and obtain a matched (fused) dataset (for more information about statistical matching, see Rodgers 1984; Kadane 2001; Moriarity and Scheuren 2001; Rässler 2004; D'Orazio et al. 2006; Moriarity 2009; Eurostat 2013).

Statistical matching can be performed at micro level or macro level. In the micro-level approach, data from individual units in the different datasets are combined to construct synthetic records containing information on all variables. Hence, a complete synthetic microdata is obtained. When statistical matching is carried out in the macro-level approach, a parametric model is first

constructed for all the data, for example, a multivariate normal model for continuous variables or a multivariate multinomial model in the case of categorical variables. Then, the parameters of this model are estimated and used to estimate population parameters of interest. We refer to D'Orazio et al. (2006) for a review of statistical matching methods at the macro level and focus here on statistical matching at the micro-level.

Statistical matching is closely related to imputation for missing data. However, imputation techniques estimate only missing values within the data, whereas statistical matching methods estimate values for all units on variables that are not observed. This is the reason why the terminology on both concepts is often used interchangeably.

Statistical matching in the micro-level approach can be applied following different methods e.g., a parametric approach (based on model predictions), a nonparametric approach (based on hot-deck techniques), and a mixed approach (combining the two). In the parametric approach, once a parametric model has been estimated, synthetic microdata are obtained as draws from the predicted distribution, using the parametric model and observed variables. To facilitate variance estimation, this procedure can be carried out multiple times. The parametric method strongly relies on the model assumptions.

With hot-deck methods, each missing value is replaced by an observed value from a similar unit with respect to background variables (McMillan 2013). If a distance (nearest neighbor) hot-deck is used, then each record in the recipient file is matched with the closest record in the donor file according to a predetermined distance measure. This technique can often be seen as a single imputation approach. Kaiser (1983) shows that distance hot-deck produces more reliable estimates than random hot-deck. Hot-deck approaches are widely applied for several reasons. First, hot deck yields realistic values since estimates use observed values from the empirical distribution and therefore cannot be outside the range of possible values. Second, it is not necessary to model the distribution of missing values (Siddique and Belin 2008; Myers 2011). However, it is crucial that good donor classes are constructed; for example, donor records may be used repeated when donor classes are too small, leading to biased variance estimates (de Waal 2015).

To benefit from properties of both parametric and nonparametric approaches, mixed methods can be used. These consist of two steps: first, a parametric model is estimated to build donor classes based on predictions from the model; second, a nonparametric approach such as a distance hot-deck is used to create the matched data. Predictive mean matching is an example of a mixed method, where the donor and recipient records have the closest predicted values according to a parametric model (Rubin 1986). Note that the parametric model approach is more parsimonious, whereas hot-deck techniques offer some protection against model misspecification. Hence, predictive mean matching is often seen as a good compromise (D'Orazio et al. 2006).

In statistical matching, the relationship between variables Y in file A and variables Z in file B cannot be estimated directly, since these variables are observed in different datasets. Indeed, in order to estimate this relationship, one needs to rely on an *untestable* assumption. The most common assumption is the conditional independence assumption (CIA), which states that conditional on the values of common variables X, the target variables Y and Z are independent. Kadane (2001), Sims (1972), and Singh et al. (1993) present studies describing the effects of the CIA on the matched file (MF) following statistical matching. However, as noted in the literature, (e.g., Cutillo and Scanu 2020), the CIA is a strong assumption.

Hence, the use of auxiliary information is helpful in the statistical matching process. For example, Singh et al. (1993) discuss how auxiliary information obtained from an additional "small" file C containing information about the correlation structure of X, Y, and Z can be included in the statistical matching procedure in order to provide a better-quality MF. The way that auxiliary information can be used varies. "Plug-in" macro-level estimates for the correlation structure can be incorporated directly into the formula for linear regression predictions as shown in this article (see also D'Orazio et al. 2006). Other micro-level approaches have adapted an EM algorithm (van Delden et al. 2020) and data augmentation (Schafer 1997; D'Orazio et al. 2006, section A.2) to obtain maximum likelihood (ML) estimates of regression parameters to be used in statistical matching similar to missing data problems. Fosdick et al. (2016) propose a statistical matching approach that allows for incorporating auxiliary information in the presence of categorical variables. Ucar and Betti (2016), interestingly, incorporate the use of auxiliary information in the case of statistical matching of longitudinal social surveys.

In this article, we investigate the effect of using auxiliary information about the joint distributions of Y, Z, and X contained in a file C on the quality of the MF under different scenarios. We review the methodology for statistical matching under the general multivariate setup and propose extending the prediction models by calibrating to known (or estimated) benchmarks in file C. The notion of calibrating model predictions is derived from research in the context of calibrated data imputation by Pannekoek et al. (2013) and Shlomo et al. (2009). We compare the mixed methods approach for statistical matching with and without the use of auxiliary information in file C, demonstrating the utility of including such available auxiliary information. Using a "robust" regression model calibrated to known or weighted population totals, we obtain a more reliable correlation structure between X, Y, and Z. This information is then "plugged" into the parametric modelling coefficients used in a predictive mean matching approach. To assess whether the calibration compensates for errors and adds robustness to the estimation of the correlation structure, we contaminate and add misspecifications to the auxiliary data in file C. The evaluation of our proposed approach focuses on the accuracy of estimated correlations and regression models parameters from the MF output, since these data

will be used for further statistical analyses in practice. We evaluate our approach in both a large-scale simulation study and on a real application.

We note here the research that has been carried out on the problem of statistical matching and imputation in the presence of complex survey designs: examples include Yang and Kim (2020), Andridge and Little (2010), Conti et al. (2017), Riddles et al. (2016), Morikawa and Kim (2018), Rodgers (1984), Rubin (1986), and Renssen (1998). These papers generally focus on incorporating the survey weights and the design variables (i.e., stratification and/or clustering variables) into the statistical matching or imputation processes. For example, Andridge and Little (2010) account for the sampling design by creating donor classes that use both design and survey variables and proposed a weighted random hot deck imputation procedure. A limitation of this approach is that the complete set of design variables may not be available, perhaps due to disclosure control measures. Although we recognize the importance of statistical matching (and more broadly imputation) under complex and informative survey designs in the social surveys context, we present our proposed approach assuming without loss of generality that the surveys have equal inclusion probability and noninformative designs. This is the case for many national surveys, such as the Labour Force Survey, Income, and Household Budget Surveys. Donatiello et al. (2018) assumed noninformative survey designs and merged the Italian European Union Statistics for Income and Living Conditions (EU-SILC) (European Union 2018) and the Italian Household Budget Survey and showed that the integrated dataset was of good quality and fit-for-purpose. We therefore adopt this framework, and the topic of more complex survey designs will be a subject of future research. That said, in our proposed approach of calibrating the prediction models in file C to known or estimated benchmarks, we implicitly account for survey weights in the files to be statistically matched.

The remainder of this article is organized as follows. In section 2, we introduce notation, provide a general framework, then present our proposed method of improving the statistical matching by calibrating the prediction models in the auxiliary file C. Section 3 presents a simulation study, comparing results obtained with CIA to those obtained using an auxiliary file C, with and without prediction model calibration. An additional model-based simulation study is presented in the supplementary data online. Section 4 presents an empirical application that uses the EU-SILC and Living Costs and Food (LCF) Survey for the United Kingdom. We conclude in section 5 with a discussion and future work.

2. STATISTICAL MATCHING APPROACH

In this section, we first introduce the notation used in this article, then describe a two-step (mixed) statistical matching technique. This method is similar to

those carried out at some NSIs (D'Orazio et al. 2005; Eurostat 2013). Finally, we present how auxiliary information in an additional file C can be used to improve the statistical matching and propose a modification for a more "robust" regression prediction model by calibrating to known (or estimated) totals.

2.1 The Framework

To introduce the statistical matching framework, we follow D'Orazio et al. (2006).

Let $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ be a random variable with density $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_P)'$, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_Q)'$, and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_R)'$ be vectors of random variables of dimension P, Q, and R, respectively. We assume that A and B are two sample datasets consisting of n_A and n_B independent and identically distributed observations generated from $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. We also assume that the units in A have $\boldsymbol{Z}$ missing and the units in B have $\boldsymbol{Y}$ missing, that is:

$$\left(\boldsymbol{x}_a^A, \boldsymbol{y}_a^A\right) = \left(x_{a1}^A, \ldots, x_{aP}^A, y_{a1}^A, \ldots, y_{aQ}^A\right),$$

$$\left(\boldsymbol{x}_b^B, \boldsymbol{z}_b^B\right) = \left(x_{b1}^B, \ldots, x_{bP}^B, z_{b1}^B, \ldots, z_{bR}^B\right),$$

where $a = 1, \ldots, n_A$ and $b = 1, \ldots, n_b$ denotes the observed values of the units in file A and B, respectively. We assume that A and B have equal probability (EPSEM) designs and are noninformative with respect to their sample designs. Furthermore, we assume that both A and B cover the same target population.

For convenience, we assume that $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ follows a multivariate normal distribution that will enable the use of linear regression models in the statistical matching, having parameters:

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XZ} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZX} & \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ} \end{pmatrix} \right],$$

with associated joint distribution denoted by $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Transformations may have been applied to the original data to obtain approximate normality. Other model extensions have been studied in the literature (see Lee and Carlin 2017; van Buuren 2018).

Under the CIA assumption, $\boldsymbol{\Sigma}_{YZ} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XZ}$. Our objective is to obtain an MF containing $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$. For this purpose, file A will be the recipient file, and file B is the donor file.

2.2 Mixed Method Approach

Here, we describe statistical matching using a two-step approach of predictive mean matching. First, an appropriate parametric model is chosen to obtain predictions. Second, a nonparametric approach employs a distance metric to find a donor record with the closest (minimum distance) prediction for each recipient. Here, we impute values $\boldsymbol{Z}$ to file A. As continuous variables such as expenditures and income are important for NSIs, we focus on continuous $\boldsymbol{Y}$ and $\boldsymbol{Z}$ variables. We use a distance hot-deck technique to match on predictions. This approach offers protection again model misspecification that can arise from the first step (D'Orazio et al. 2006). In case of continuous variables, this approach is well studied and adopted in the literature. Transformations can be used to apply the models in case of skewed variables (D'Orazio et al. 2006). For ordinal or categorical variables, other models need to be considered, such as multinomial distributions (e.g., de Waal 2015; Takabe and Yamashita 2020).

2.2.1 Assuming the CIA

The CIA posits that the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ can be explained entirely by the values of the variables $\boldsymbol{X}$ (de Waal 2015).

As a first step under this assumption, the following regression models are estimated on files A and B, respectively:

$$\boldsymbol{Y} = \boldsymbol{\alpha}_{\boldsymbol{Y}} + \boldsymbol{\beta}_{\boldsymbol{YX}}\boldsymbol{X} + \boldsymbol{e}_{\boldsymbol{Y}|\boldsymbol{X}}, \tag{1}$$

$$\boldsymbol{Z} = \boldsymbol{\alpha}_{\boldsymbol{Z}} + \boldsymbol{\beta}_{\boldsymbol{ZX}}\boldsymbol{X} + \boldsymbol{e}_{\boldsymbol{Z}|\boldsymbol{X}} \tag{2}$$

with $\boldsymbol{e}_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\boldsymbol{e}_{\boldsymbol{Z}|\boldsymbol{X}}$ distributed as a multivariate normal with null mean vectors and covariance matrices denoted by $\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{X}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{Z}|\boldsymbol{X}}$.

Once these regression model parameters are estimated, the statistical matching is carried out as follows:

(1) ***Regression step.*** Compute "intermediate values" for units a in A and b in B as:

$$\tilde{\mathbf{z}}_a = \hat{\boldsymbol{\alpha}}_{\boldsymbol{Z}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{ZX}}\boldsymbol{x}_a, \tag{3}$$

$$\tilde{\mathbf{y}}_b = \hat{\boldsymbol{\alpha}}_{\boldsymbol{Y}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{YX}}\boldsymbol{x}_b, \tag{4}$$

where $\hat{\boldsymbol{\alpha}}_{\boldsymbol{Z}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{Z}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\,\hat{\boldsymbol{\mu}}_{\boldsymbol{X}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{ZX}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}$, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{Y}} = \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}\,\hat{\boldsymbol{\mu}}_{\boldsymbol{X}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{YX}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1}$ are the ML estimators of $\boldsymbol{\alpha}_{\boldsymbol{Z}}$, $\boldsymbol{\beta}_{\boldsymbol{ZX}}$, $\boldsymbol{\alpha}_{\boldsymbol{Y}}$, $\boldsymbol{\beta}_{\boldsymbol{YX}}$, respectively.

(2) ***Matching step.*** For each $a = 1, \ldots, n_A$ in A, impute z_{b^*} corresponding to the nearest neighbor b^* in B with respect to a constrained distance function $d((\boldsymbol{y}_a, \tilde{\mathbf{z}}_a), (\tilde{\mathbf{y}}_b, \mathbf{z}_b))$. Here, we use the Manhattan distance to be consistent with de Waal (2015), Linskens (2015), and van Roij (2015). Note that de Waal et al. (2017) evaluated a variety of different distance measures for

hot deck, with similar results for all distance measures. Rodgers (1984) proposed the Manhattan distance as it provided an MF of good quality and was particularly robust in the presence of outliers in numerical data.

2.2.2 Including auxiliary information from an additional file C

When the CIA does not hold, the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ cannot be estimated from files A and B. The CIA is a strong assumption (Cutillo and Scanu 2020) that does not generally hold in practice. Therefore, the use of auxiliary information is crucial. Singh et al. (1993) recommend using auxiliary information to specify a plausible relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. For example, this relationship can be approximated by a correlation structure between the variables estimated from an additional "small" file C wherein the variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ are jointly observed, as in this paper.

This approach involves the same steps as outlined above in section 2.2.1 but with a modification of the predictions in step 1 as follows:

$$\tilde{\mathbf{z}}_a = \hat{\boldsymbol{\mu}}_{\boldsymbol{Z}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}|\boldsymbol{Y}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{XX}|\boldsymbol{Y}}(\boldsymbol{x}_a - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}) + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZY}|\boldsymbol{X}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{YY}|\boldsymbol{X}}(\boldsymbol{y}_a - \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}}), \tag{5}$$

$$\tilde{\mathbf{y}}_b = \hat{\boldsymbol{\mu}}_{\boldsymbol{Y}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}|\boldsymbol{Z}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{XX}|\boldsymbol{Z}}(\boldsymbol{x}_b - \hat{\boldsymbol{\mu}}_{\boldsymbol{X}}) + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZY}|\boldsymbol{X}}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{ZZ}|\boldsymbol{X}}(z_b - \hat{\boldsymbol{\mu}}_{\boldsymbol{Z}}). \tag{6}$$

As highlighted in D'Orazio et al. (2006), when using auxiliary information, the modelling setting is identical to the one described in (1) and (2). The additional information of the correlation structure affects *only* the prediction estimation phase as shown in (5) and (6). The unknown parameters in (5) and (6) are estimated via ML, which is well supported by the literature in the statistical matching context. The interested reader may want to refer to D'Orazio et al. (2005) where an extensive simulation study is carried out to compare different estimation techniques. We refer to D'Orazio et al. (2006) and Moriarity and Scheuren (2001) for the derivations of the estimators.

2.2.3 Calibrating prediction models in file C

The file C that contains auxiliary information about the joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ may suffer from various issues. For example, file C might be outdated, contain proxy variables, or contain variables drawn from different distributions from their counterparts in files A and B. Therefore, it is important to consider these potential misspecifications when the relationships between the variables in file C are used in the statistical modelling in step (1) of section 2.2.2.

Here, we consider a calibration strategy on file C to add "robustness" to the model for estimating the covariance matrices used in predictions (5) and (6). Our hypothesis is that this modification allows us to provide better-quality estimates of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZY}|\boldsymbol{X}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX}|\boldsymbol{Z}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX}|\boldsymbol{Y}}$ hence protecting from potential model misspecification.

In particular, the regression models in file C are transformed, and an extra row and two extra columns are added to the design matrix as follows:

$$\begin{pmatrix} \boldsymbol{X}_1 & \dots & \boldsymbol{X_P} & \boldsymbol{Y}_1 & \dots & \boldsymbol{Y_Q} & \boldsymbol{Z}_1 & \dots & \boldsymbol{Z_R} & \boldsymbol{r} & \boldsymbol{t} \\ X_{1t}-\hat{X}_{1t,s} & \dots & X_{Pt}-\hat{X}_{Pt,s} & Y_{1t}-\hat{Y}_{1t,s} & \dots & Y_{Qt}-\hat{Y}_{Qt,s} & Z_{1t}-\hat{Z}_{1t,s} & \dots & Z_{Rt}-\hat{Z}_{Rt,s} & N-n_C & 0 \end{pmatrix} \quad (7)$$

The last row in (7) contains the differences between external population totals and the weighted sample totals of the relevant variables in file C. Note that the subscript 't' denotes that the variable is an external total and 't,s' denotes the total estimated from file C. The external population totals can be taken from a Census, administrative data or a large survey sample, in this case the estimated weighted totals from the recipient file A to be statistical matched. In addition, $\boldsymbol{r} = (0, 0, \dots, 0)'$, and $\boldsymbol{t} = (1, 1, \dots, 1)'$.

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZY|X}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{YX|Z}}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{ZX|Y}}$ are obtained from the calibrated regression models from file C and used in the statistical matching as described in sections 2.2.1 and 2.2.2.

3. SIMULATION STUDIES

We present two simulation studies designed to evaluate the performance of the approaches outlined in section 2. The setting of the first simulation assumes that the original population is distributed as multivariate normal c.f. D'Orazio et al. (2005). The second simulation study investigates sensitivity to model misspecification when the population is not distributed as multivariate normal. To save space, the second simulation is included in the supplementary data online.

Three strategies are evaluated:

- **CIA:** Assuming conditional independence (section 2.2.1),
- **Auxiliary *C*:** Including auxiliary information from an additional file C (section 2.2.2), and
- **Auxiliary Calibrated *C*:** Including auxiliary information through calibrated regression models in file C (section 2.2.3).

3.1 Generating the Population and Simulation Steps

Let (X_1, X_2, Y, Z) each be univariate standard normally distributed variables with joint distribution density $f(X_1, X_2, Y, Z) \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{X_1} & & & \\ \rho_{X_1X_2}\sqrt{\sigma^2_{X_1}\sigma^2_{X_2}} & \sigma^2_{X_2} & & \\ \rho_{X_1Y}\sqrt{\sigma^2_{X_1}\sigma^2_{Y}} & \rho_{X_2Y}\sqrt{\sigma^2_{X_2}\sigma^2_{Y}} & \sigma^2_{Y} & \\ \rho_{X_1Z}\sqrt{\sigma^2_{X_1}\sigma^2_{Z}} & \rho_{X_2Z}\sqrt{\sigma^2_{X_2}\sigma^2_{Z}} & \rho_{YZ}\sqrt{\sigma^2_{Y}\sigma^2_{Z}} & \sigma^2_{Z} \end{bmatrix},$$

with $\boldsymbol{\mu} = (10,\ 15,\ 9,\ 11)'$ and $\sigma^2_{X_1} = \sigma^2_{X_2} = \sigma^2_Y = \sigma^2_Z = 1$. The population size is $N = 50{,}000$. In our simulation study, different levels of correlations are chosen: $\rho_{X_1X_2} = 0.001$, $\rho_{X_1Y} = 0.20$, $\rho_{X_1Z} = 0.50$, $\rho_{X_2Y} = 0.20$, $\rho_{X_2Z} = 0.50$, and $\rho_{YZ} = \{0.01, 0.20, 0.50, 0.75\}$. The simulation involves the following steps, repeated independently 500 times:

(1) *Sampling*: From the fixed population, we draw three independent samples A, B, and C using simple random sampling without replacement with sample sizes $n_A = 2{,}000$, $n_B = 3{,}000$, and $n_c = 250$.
We assume the following: $A = (X_1, X_2, Y)$ and $B = (X_1, X_2, Z)$, where A is the recipient file and B the donor file, and $C = (X_1, X_2, Y, Z)$. Sample C is used as auxiliary information. In this case, sample C is not contaminated and comes from the same distribution.
(2) *Contamination of sample C*: We contaminate Y and Z in sample C, simulating inconsistency in variable measurement. In particular, we consider a contamination scenario where we add noise $\tau_c \sim N(0,\ 0.1)$ to 10% of records of file C.
(3) *Statistical matching step*: Following the strategies described in section 2, match the donor file B to the recipient file A to obtain a MF.
(4) *Evaluations*: To evaluate the results and compare the alternative approaches, we calculate several statistics from each of the MFs generated in the simulation, evaluating the properties of each statistic over repeated samples using the absolute relative bias (RAB) defined in (8):

- pairwise correlation of all variables, focusing on ρ_{YZ};
- simulation variance of each variable in the MF;
- regression parameters for $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_1 + \beta_3 x_2 + e_i,\ \ i \in \text{MF}$.

3.2 Results

Figures 1 through 3 present the RAB where $\hat{\theta}_i$ is the statistic of interest computed from iteration s ($s = 1,\ldots 500$) and θ is the true value obtained from the fixed population generated as described in section 3.1 calculated as:

$$\text{RAB} = \frac{1}{500} \sum_{s=1}^{500} \left| \frac{\hat{\theta}_s - \theta}{\theta} \right| \tag{8}$$

Figure 1 presents bar charts of the RAB for the correlation between Y and $Z(\rho_{YZ})$ obtained with each statistical matching strategy for each correlation level, ($\rho_{YZ} = \{0.01, 0.20, 0.50, 0.75\}$), with uncontaminated auxiliary data (C) and contaminated auxiliary data. Correlations of variables originating from the same (or recipient) file have low RAB, largely explained by sampling error. When $\rho_{YZ} = 0.01$, there is little improvement in precision by including file C, although the calibration strategy offers slight improvment. However, when file C is contaminated with additive noise, the calibration appears to reduce the effects of model misspecification, leading to less biased correlation estimates. We see a similar pattern when ρ_{YZ} increases to 0.20. The CIA and calibrated C perform similarly under the correct model, with both yielding less biased estimates than those obtained using file C without the calibration step. We also see a smaller bias from implementing the calibration step when C is contaminated with noise. When ρ_{YZ} increases to 0.50 and 0.75, using file C results in less biased correlation estimates than those obtained with CIA. When file C is used with and without the noise contamination, there is little impact from the calibration step for $\rho_{YZ} = 0.5$, although the calibration appears to increase the RAB when $\rho_{YZ} = 0.75$. In summary, figure 1 provides evidence that the Auxiliary C procedure yields less biased correlation estimates than the CIA-counterparts when ρ_{YZ} is "relatively high" and that calibration of the model in C helps in most cases, particularly when ρ_{YZ} is small.

Figure 2 presents bar charts of the variance of $Z(\boldsymbol{\sigma}_z^2)$ in the MFs obtained with each statistical matching strategy for each correlation level. RAB is compared to the true value $\boldsymbol{\sigma}_z^2 = 1$. Recall that the value of Z was taken from the donor file B and attached to file A. Therefore, it is of interest to see if $\boldsymbol{\sigma}_z^2$ is preserved in the MF. We note that it was possible to have multiple donors from file B for each recipient in file A, although this would be unlikely given the larger pool of donors ($n_A = 2{,}000$ and $n_B = 3{,}000$). Again, evaluations of univariate statistics for other variables in the MFs have low RAB, largely explained by sampling error. Figure 2 presents strong evidence that the Auxiliary Calibration C procedure substantively reduces the bias of the estimated σ_z^2 for all levels of ρ_{YZ}, even if the auxiliary file is contaminated.

Finally, recall that we fit the following regression model in each candidate MF: $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_1 + \beta_3 x_2 + e_i,\ i \in \text{MF}$. We focus on the coefficient for the variable Z (β_1). Figure 3 presents the RAB of β_1 for the uncontaminated and contaminated with noise settings. When ρ_{YZ} is small, using auxiliary information in file C offers little improvement over the CIA. This changes when ρ_{YZ} is large ($\rho_{YZ} = 0.75$). In all instances, however, the Auxiliary Calibrated C procedure produces less biased estimates of β_1 than those obtained via other statistical matching procedures, although it should be noted

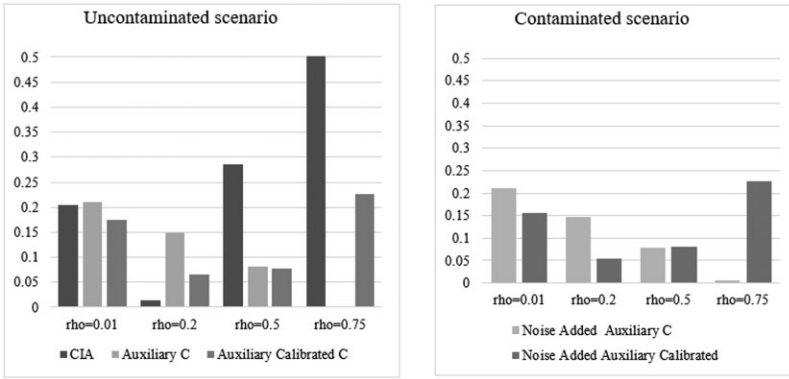

Figure 1. Relative Absolute Bias (RAB) of Correlations between *Y* and *Z*(ρ_{YZ}): Uncontaminated File *C* (Left) and Contaminated with Noise Addition in File *C* (Right).

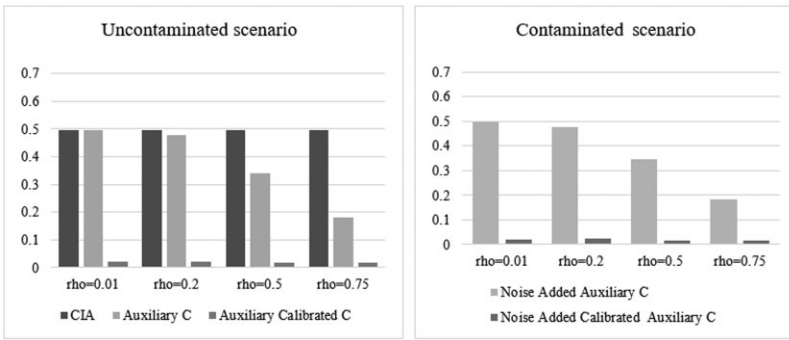

Figure 2. Relative Absolute Bias (RAB) of Variance of Z (true value of $\sigma_z^2 = 1$): Uncontaminated File *C* (Left) and Contaminated with Noise Addition in File *C* (Right).

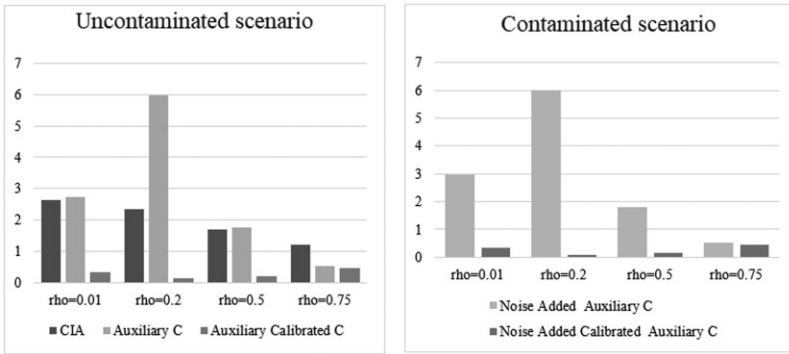

Figure 3. Relative Absolute Bias (RAB) of Regression Coefficient β_1 for the Regression Model: $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_1 + \beta_3 x_2 + e_i, i \in$ MF: Uncontaminated File *C* (Left) and Contaminated with Noise Addition in File *C* (Right). True values are $\beta_1 = -0.381$ for $\rho_{YZ} = 0.01$; $\beta_1 = -0.002$ for $\rho_{YZ} = 0.20$; $\beta_1 = 0.595$ for $\rho_{YZ} = 0.50$; and $\beta_1 = 1.100$ for $\rho_{YZ} = 0.75$.

that the level of RAB still remains high when $\rho_{YZ} = 0.75$ compared to other levels of ρ_{YZ}. Importantly, even when file *C* is perturbed by adding noise to 10 percent of the records, the RAB is smaller when carrying out the calibration on file *C* for all levels of ρ_{YZ}.

3.3 Final Remarks on the Simulation Studies

The simulation study provides evidence that our proposed calibration approach on the prediction models in file *C* improved the quality of the MF over the other considered statistical matching methods. This approach consistently yielded less biased variance estimates of Z (the variable that is attached to file *A* from file *B*), as well as less biased linear regression slope parameters for the fitted model. The simulation study presented in the appendix in the supplementary data online provides evidence that the calibration of the prediction models on file *C* helps mitigate model failures such as nonnormality of error terms and heteroscedasticity. Thus, if primary usage of the MF is regression analyses, we advise using an auxiliary file *C* with calibration to known or estimated totals to improve the quality of the statistical matching.

4. APPLICATION

In this section, we studied the statistical matching strategies to empirical UK data in an application designed to study the relationship between housing deprivation and household expenditure and other socio-demographic variables. We match the EU-SILC 2018 (European Union 2018) to the LCF Survey 2017–2018 (Office for National Statistics, Department for Environment, Food and Rural Affairs 2020). We follow the approach outlined in section 2 as it forms the basis of data integration of social surveys in many NSIs (see Eurostat 2013; Donatiello et al. 2018). We assume an auxiliary file *C* and demonstrate our proposed approach of calibrating the prediction models in file *C* to improve the quality of the statistical matching.

4.1 The Data

The EU-SILC is an important survey aiming at collecting cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. The LCF collects information on regular expenditures, such as rent, food, and mortgage payments on all persons aged 16 years and over in the household via a household interview. Both surveys contain files at both household and personal levels. Expenditure data are collected in the LCF and are not collected in the EU-SILC. Personal income is collected in both surveys. Therefore, our goal is to create a single dataset containing

deprivation, social exclusion, income, and expenditure variables. In this study, the EU-SILC is the recipient file, and the LCF is the donor file. In the UK, both surveys have the same target population and have similar patterns of nonresponse. In addition, both surveys have the same two-stage design with equal final inclusion probabilities (EPSEM designs). We therefore do not incorporate the survey weights directly in this application but note that calibrating to weighted survey counts in the auxiliary file as outlined in section 2.2.3 indirectly accounts for differential survey weights. Incorporating the survey weights more directly is a subject of future research. Before applying the statistical matching, the matching variables are harmonized across datasets. First, we attach variables from the household file to the personal file of the EU-SILC to obtain the deprivation variables used in the application. Second, the statistical matching procedures described in section 2.2 are applied at person level to obtain MFs. Finally, we revert each MF back to a household file by considering the EU-SILC household reference person to perform further statistical analysis on deprivation. Since 2001/2002, the same concept of household reference person has been used on all the UK Government surveys (Eurostat 2013; Office for National Statistics 2021), hence both surveys are consistent in this sense. The sample sizes of the EU-SILC and LCF personal files are $n = 17,645$ and $n = 12,753$, respectively, whereas the sample sizes of the EU-SILC and the LCF household files are $n = 9,711$ and $n = 5,407$, respectively.

4.2 Statistical Matching

In this application, we apply the statistical matching strategies described in section 2.2. The first strategy assumes CIA, the second strategy incorporates the correlation structure between the total personal expenditure and personal income as shown with an auxiliary file C, and the third strategy incorporates the correlation structure after calibrating the model in the auxiliary file C. In this case, however, the total personal expenditure and personal income are also available in the LCF. Therefore, this application modifies the approach presented in section 2 as the LCF can be seen as both a file B and a file C. Consequently, the calibration of the model in file C is carried out according to the weighted survey totals of file A (EU-SILC) for the X variables and total personal income, but there is no need to calibrate the expenditure variables in file B since file B and file C are the same file. The regression models used to obtain predictions are given below, with income and expenditures variables log transformed in all applications as is customary.

Assuming CIA, the following regression model is fit on the LCF:

Table 1. Regression Parameters for Model (9) from the LCF

Variable	Estimate
Intercept	4.108*
Age	0.006*
Male	−0.292*
Married	0.010
Separated	0.034
Widowed	−0.382*
Divorced	−0.106
Unemployed	−0.622*
Student	−0.983*
Retired	−0.271*
Good health	0.529*

NOTE.—*indicates significant (p-value $<.05$).

$$\begin{aligned}\log(\text{TotalPersonalExpenditure}_i) = {} & \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Age}_i + \beta_3 \text{Married}_i \\ & + \beta_4 \text{Separated}_i + \beta_5 \text{Widowed}_i \\ & + \beta_6 \text{Divorced} + \beta_7 \text{Unemployed} \\ & + \beta_8 \text{Student} + \beta_9 \text{Retired} \\ & + \beta_{10} \text{GoodHealth} + e_i \end{aligned} \tag{9}$$

with $e_i \sim N(0,\ \sigma_e^2)$ and i.i.d. See table 1.

Using the realized regression model parameter estimates from (9), the predictions of total personal expenditure are obtained in EU-SILC.

Next, we apply distance hot-deck with the Manhattan distance function with the following matching variables: log of the total personal expenditure and log of the total personal gross income. Personal gross income is available (hence not predicted) on both datasets.

A slight modification of the second strategy that makes use of the auxiliary file *C* is required, as the LCF (which serves as files *B* and *C*) contains total personal expenditures and total personal gross income. The correlation between the total personal expenditure and total personal gross income in the LCF is equal to $\hat{\rho} = 0.34$. We proceed as follows:

(1) Estimate the correlation between the total personal expenditure and total personal gross income in the LCF (logarithm transformed to attenuate the skewness).
(2) Estimate model (9) in the LCF as above.
(3) Estimate the following regression model in the EU-SILC as:

Table 2. Regression Parameters for Model (10) from the EU-SILC

Variable	Estimate
Intercept	2.068*
Age	0.031*
Male	0.225*
Married	−0.424*
Separated	−0.209
Widowed	0.013
Divorced	−0.115
Unemployed	−0.375*
Student	−0.840*
Retired	2.701*
Good health	3.596*

NOTE.—*indicates significant (p-value $<.05$).

$$\begin{aligned}\log(\text{TotalPersonalGrossIncome}_i) = {} & \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Age}_i + \beta_3 \text{Married}_i \\ & + \beta_4 \text{Separated}_i + \beta_5 \text{Widowed}_i \\ & + \beta_6 \text{Divorced} + \beta_7 \text{Unemployed} \\ & + \beta_8 \text{Student} + \beta_9 \text{Retired} \\ & + \beta_{10} \text{GoodHealth} + e_i, \end{aligned} \tag{10}$$

with $e_i \sim N(0,\ \sigma_e^2)$ and i.i.d. See table 2.

(4) Apply the statistical matching approach described in section 2.2.2, where the new predictions from (5) and (6) take into account the estimated correlation obtained above in step (1).

The donor is chosen via the distance hot-deck algorithm with respect to a constrained Manhattan distance function calculated on the model predictions.

To apply the third strategy that calibrates the prediction model in file C (which in this case is file B), we use the same auxiliary variables as used in model (10). The benchmarks for calibration are obtained from the weighted survey totals of file A (EU-SILC) for the auxiliary variables X and personal income.

4.3 Analysis

The first evaluation of the MF considers the following regression model:

$$\begin{aligned}\log(\text{TotalExpenditure}_i) = {} & \beta_0 + \beta_1 \text{Male}_i + \beta_2 \text{Age}_i + \beta_3 \text{Married}_i \\ & + \beta_4 \text{Separated}_i + \beta_5 \text{Widowed}_i + \beta_6 \text{Divorced} \\ & + \beta_7 \text{Unemployed} + \beta_8 \text{Student} + \beta_9 \text{Retired} \\ & + \beta_{10} \text{GoodHealth} + \beta_{11} \text{Income} + e_i, \end{aligned} \tag{11}$$

with $e_i \sim N(0, \ \sigma_e^2)$ and i.i.d.

This model is first fit on the LCF survey, which contains *all* variables (including income). Then, (11) is fit to the MF under CIA, to the MF where auxiliary information is used according to section 2.2.2 (Auxiliary *C*), and to the MF obtained with the calibration step in file *C* according to section 2.2.3 (Auxiliary Calibrated *C*). The personal income variable in the MF is obtained from the EU-SILC. In this way, we are able to examine the quality of the MF in terms of preserving relationships between variables. Table 3 presents the results of the regression model in (11) on the original LCF with income taken from the LCF and for the three statistical matching scenarios with income taken from EU-SILC.

Table 3 shows that the use of auxiliary information in file *C* helps preserve the relationship between expenditure and income in the resultant MF. On the whole, the regression coefficients obtained from the Auxiliary Calibration *C* procedure tend to be the most similar to their (original) LCF counterparts.

We next describe our more complex application in the matched dataset where measures of deprivation are obtained from EU-SILC data. As a first step, we produce a multidimensional housing quality indicator on the MF following Moretti et al. (2020) latent dimensions. In particular, a two-factor confirmatory factor analysis (CFA) model is estimated with two dimensions, i.e., Housing Material Deprivation and Residential Area Deprivation (Moretti et al. 2020). The variables assigned to each latent dimension are as follows:

- Housing Material Deprivation: log income, severe material deprivation, housing ownership and problems in the house (e.g. darkness); and
- Residential Area Deprivation: pollution, grime or other environment problems, crime, violence or vandalism in the area, and noise from neighbors or from the street.

The factor scores are then estimated and used as response variables in the analysis described below.

The CFA model performs well, with the following goodness of fit indices: RMSE = 0.051, SRMR = 0.039 and CFI = 0.930 (Hu and Bentler 1999). The two latent variables produced are highly correlated with a correlation coefficient equal to 0.43. Table 4 shows the results of two regression models estimated on the each MF obtained with the considered statistical matching procedures where the response variables are the factor scores (on log scale) of housing material deprivation and residential area deprivation, respectively. The

expenditure covariates in the regression models have been scaled by dividing by 1,000.

By comparing the regression modelling results in table 4, we see some differences on the coefficient estimates between the approaches of CIA and the use of auxiliary information (with and without calibration). Recall that when the correlation between Y and Z variables is approximately equal to 0.5, the use of auxiliary information and the calibration approach in file C improves the performance of the statistical matching. In this application, the correlation is equal to 0.34, approaching that condition. We therefore argue that the use of auxiliary information and calibration in file C as shown in sections 2.2.2 and 2.2.3 should improve the regression modelling results in the MF.

Table 4 shows several consistent patterns, regardless of statistical matching procedure. As people age, they are less deprived in both dimensions. Unemployed people are more deprived on average than employed people, and retired people are in turn less deprived than employed people. Males are less deprived than females on average. People classified as Asian are less deprived than those classified as White. However, there are mixed results for people in the Black ethnic group category where there is an interchange of more and less deprivation across the two dimensions. In line with the deprivation literature (see, e.g., de Noronha 2015), the people in the Black ethnic group show higher levels of housing material deprivation compared to White ethnic group people when computed from the MFs that utilize auxiliary information (Auxiliary C and Auxiliary Calibrated C). On the other hand, incorporating the auxiliary information in the matching process tends to reduce residential area deprivation and perhaps inflate residential area deprivation in the Other ethnic group. We note that the ethnicity variable is not present in the EU-SILC data and is attached to the matched data from the LCF as a result of the statistical matching process. In addition, it is possible that these broad ethnic groups in the application have more heterogeneity leading to mixed results.

Of more interest is the relationship between expenditures and deprivation in table 4 as it can only be observed through the statistical matching application. There is a positive relationship between water and electricity expenditure in both deprivation dimensions in the case of statistical matching under CIA. However, the relationship becomes negative in the case of the MF obtained by using the auxiliary information. As evidenced in the literature, we would expect to see this negative relationship (see, e.g., Deutsch et al. 2015), supporting the outcomes in the latter. For the housing material deprivation, there is a negative relationship for the health expenditure. There is a mixed result for the residential area deprivation. Deutsch et al. (2013) argue that when people become deprived, health expenditures (such as dental expenditures) are cut back or suppressed, and we would expect to see the negative relationship. We note that those that have higher expenditures on furniture seem to have higher deprivation in the two dimensions studied which seems to be a surprising result and deserving of more investigation in future research.

Table 3. Results of Regression Model Given in (11) on the Original LCF and the Statistically Matched Files According to CIA, Using Auxiliary Information in File *C* (Auxiliary *C*) and Calibrating the Prediction Model in File *C* (Auxiliary Calibrated *C*). Estimated regression parameters are presented with 95-percent confidence intervals.

Variable	Original LCF	Statistically matched files		
		CIA	Auxiliary *C*	Auxiliary Calibrated *C*
Intercept	3.094 (2.921, 3.267)	4.080 (4.072, 4.088)	3.749 (3.739, 3.759)	3.550 (3.400, 3.611)
Age	0.004 (0.002, 0.006)	0.005 (0.005, 0.005)	0.001 (0.001, 0.001)	0.003 (0.001, 0.004)
Sex	−0.383 (−0.430, −0.336)	−0.241 (−0.244, −0.238)	−0.294 (−0.298, −0.290)	−0.300 (−0.350, −0.299)
Married	0.033 (−0.038, 0.104)	0.010 (0.006, 0.014)	0.040 (0.042, 0.052)	0.029 (0.020, 0.035)
Separated	0.028 (−0.088, 0.144)	0.034 (0.026, 0.042)	0.047 (0.037, 0.058)	0.030 (0.024, 0.035)
Widowed	−0.343 (−0.427, −0.259)	−0.367 (−0.373, −0.360)	−0.394 (−0.402, −0.385)	−0.333 (−0.336, −0.299)
Divorced	−0.129 (−0.256, −0.003)	−0.098 (−0.103, −0.092)	−0.080 (−0.087, −0.073)	−0.089 (−0.092, −0.086)
Unemployed	−0.116 (−0.279, 0.048)	−0.563 (−0.574, −0.553)	−0.298 (−0.311, −0.294)	−0.201 (−0.284, −0.198)
Student	−0.307 (−0.447, −0.168)	−0.709 (−0.725, −0.694)	−0.541 (−0.561, −0.521)	−0.243 (−0.347, −0.210)
Retired	−0.166 (−0.244, −0.087)	−0.241 (−0.247, −0.235)	−0.576 (−0.584, −0.568)	−0.199 (−0.300, −0.210)
Good/fair health	0.453 (0.361, 0.544)	0.571 (0.566, 0.576)	0.071 (0.065, 0.077)	0.333 (0.200, 0.401)
Income	0.164 (0.148, 0.180)	−0.002 (−0.003, −0.001)	0.143 (0.143, 0.144)	0.155 (0.144, 0.169)

Table 4. Log Housing Material Deprivation Regression Model and Log Residential Area Deprivation Regression Model Results According to CIA, Using Auxiliary Information in File *C* (Auxiliary *C*) and Calibrating the Models in File *C* (Auxiliary Calibrated *C*). Estimated regression parameters are presented with 95-percent confidence intervals.

Variable	Log housing material deprivation			Log residential area deprivation		
	CIA	Auxiliary *C*	Auxiliary Calibrated *C*	CIA	Auxiliary *C*	Auxiliary Calibrated *C*
Intercept	−0.950 (−0.990, −0.909)	−1.034 (−1.077, −0.990)	−0.922 (−0.982, −0.862)	−2.203 (−2.274, −2.131)	−2.202 (−2.278, −2.128)	−2.133 (−2.325, 2.258)
Age	−0.008 (−0.009, −0.008)	−0.005 (−0.006, −0.005)	−0.006 (−0.007, −0.005)	−0.008 (−0.009, −0.007)	−0.007 (−0.008, −0.006)	−0.007 (−0.008, 0.005)
Unemployed (employed reference)	0.135 (0.021, 0.249)	0.015 (−0.076, 0.107)	0.098 (0.033, 0.110)	0.111 (−0.088, 0.310)	0.096 (−0.062, 0.253)	0.096 (−0.063, 0.269)
Retired	−0.031 (−0.030, 0.031)	−0.031 (−0.058, −0.003)	−0.018 (−0.040, 0,010)	−0.022 (−0.075, 0.031)	−0.059 (−0.108, −0.011)	−0.025 (−0.118, −0.019)
Sex (male reference)	−0.001 (−0.022, 0.020)	−0.056 (−0.079, −0.033)	−0.014 (−0.020, 0.246)	−0.010 (−0.046, 0.027)	−0.042 (−0.082, −0.002)	−0.017 (−0.090, −0.008)
Ethnicity: mixed (white reference)	−0.025 (−0.120, 0.069)	0.058 (−0.041, 0.156)	0.018 (−0.005, 0.029)	−0.079 (−0.245, 0.086)	0.068 (−0.102, 0.239)	0.029 (−0.001, 0.35)
Asian/Asian British	−0.082 (−0.117, −0.047)	−0.084 (−0.120, −0.048)	−0.104 (−0.165, −0.044)	−0.084 (−0.145, −0.024)	−0.086 (−0.149, −0.024)	−0.085 (−0.169, −0.028)
Black/Black British	−0.085 (−0.148, −0.021)	0.029 (−0.064, 0.121)	0.082 (−0.194, 0.071)	0.017 (−0.094, 0.127)	−0.056 (−0.216, 0.103)	−0.069 (−0.114, 0.108)
Other	0.125 (0.032, 0.217)	0.163 (0.043, 0.283)	0.112 (0.050,0.123)	0.148 (−0.013, 0.309)	0.022 (−0.185, 0.230)	0.155 (−0.141, 0.100)
Water and electricity expenditure[a]	0.858 (0.758, 0.958)	−0.184 (−0.322, −0.045)	−0.100 (−0.175, −0.095)	0.358 (0.183, 0.532)	−0.210 (−0.449, 0.029)	−0.100 (−0.100, 0.014)
Health expenditure[a]	−0.102 (−1.216, 1.011)	−1.714 (−2.663, −0.765)	−0.987 (−1.690, −0.647)	0.182 (−1.762, 2.125)	−1.070 (−2.708, 0.569)	−0.988 (−1.608, 0.461)
Furniture expenditure[a]	0.235 (−0.009, 0.478)	0.378 (0.141, 0.615)	0.101 (0.017, 0.152)	0.112 (−0.313, 0.537)	0.021 (−0.388, 0.431)	0.082 (−0.245, 0.231)

NOTE.—[a]Divided by 1,000.

5. CONCLUSION

In this article, we examined a mixed method two-step approach to statistical matching. In general, the relationship between Y and Z variables cannot be estimated directly since these are observed in files A and B separately. The circumvent this, practitioners often assume conditional independence (CIA) of Y and Z, given common variables X. However, this is a very strong assumption and not likely held in practice. We therefore assess ways of relaxing this assumption using an additional auxiliary file C. To add robustness to the estimation of the correlation structure of the variables of interest, we proposed calibrating the prediction models in file C. Robustness is desirable as file C might be contaminated. We find that calibrating the prediction models within the modelling framework helps mitigate against the impact of model-misspecification and model failures and therefore improves the quality of the matched data. Since researchers will use the MF for further statistical analysis, such as regression analysis, it is crucial to provide good quality estimates of the relationships between variables.

Our simulation studies show that when the correlation between Y and Z is very small, in our case equal to 0.01, the use of additional information obtained from file C does not improve greatly the results in estimating the correlation. However, results improve considerably when this correlation increases. We demonstrated further improvements with calibration of the prediction models in file C particularly when the distributions of variables in file C are contaminated by random noise. The use of calibrating prediction models in file C also provided more consistent estimates of variances and better-quality regression model parameter estimates. In addition, the simulation study in the supplementary data online showed that calibrating the prediction models in file C helped mitigate model failures such as non-normality and heteroscedastic settings.

The empirical application in section 4 that assessed factors that impact two dimensions of deprivation provided results consistent with the literature. Of particular interest was to investigate the factors related to expenditures and their effect on the two dimensions of deprivation as this is rarely studied in the literature due to the lack of a single data source containing this combination of variables; this relationship could be studied after statistical matching. Although some of the coefficients on the expenditures were not significant, we did see a surprising outcome that higher furniture expenditure showed higher deprivation. Other expenditures on health and utilities showed a more expected relationship with the two dimensions of deprivation. Regarding the levels of significance and statistical testing in our application, we note that more work is needed on estimating variances of estimated parameters when statistical matching has been applied since naive using the current variance estimates do not account for the uncertainty arising from the statistical matching. This important is a subject of future research.

Future work will evaluate other types of contamination settings and model failures in file *C* and investigate other types of correction techniques, for example compensating directly for measurement error in the prediction models. In addition, other modelling strategies can be developed and applied in the statistical matching context, particularly focusing on incorporating survey weights and other design variables more directly when survey designs are informative.

Based on the research shown here, our recommendation is that practitioners should use auxiliary information when the correlations between the variables that are not jointly observed is expected to be medium or large. File *C* should be calibrated where population totals are available (either known or estimated from large surveys), in order to protect against potential model misspecification and model errors. This helps in providing better-quality regression model coefficient estimates in the MF, which is a crucial issue since MFs are typically used in practice for secondary data analysis.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

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

Conti, P. L., Marella, D., and Scanu, M. (2017), "Statistical Matching Analysis for Complex Survey Data with Applications," *Journal of the American Statistical Association*, 516, 1715–1725.

Cutillo, A., and Scanu, M. (2020), "A Mixed Approach for Data Fusion of HBS and SILC," *Social Indicators Research*, 150, 411–437.

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

———. (2005), "A Comparison among Different Estimators of Regression Parameters on Statistically Matched Files through an Extensive Simulation Study," Technical Report, Contributi 2005/10, Istituto Nazionale di Statistica, Rome.

de Waal, T. (2015), "Statistical Matching: Experimental Results and Future Research Questions," Discussion Paper of CBS, Statistics Netherlands.

de Waal, T., Coutinho, W., and Shlomo, N. (2017), "Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions," *Journal of Survey Statistics and Methodology*, 5, 372–397.

de Noronha, N. (2015), "Ethnic Disadvantage in the Housing Market: Evidence from the 2011 Census," A Race Equality Foundation Briefing Paper. Race Equality Foundation. Available at https://raceequalityfoundation.org.uk/wp-content/uploads/2018/02/Housing-Briefing-26.pdf.

Deutsch, J., Lazar, A., and Silber, J. (2013), "Becoming Poor and the Cutback in the Demand for Health Services in Israel," *Israel Journal of Health Policy Research*, 2, 49.

Deutsch, J., Guio, A.-C., Pomati, M., and Silber, J. (2015), "Material Deprivation in Europe: Which Expenditures Are Curtailed First?," *Social Indicators Research*, 120, 723–740.

Donatiello, G., D'Orazio, M., Frattarola, D., Scanu, M., and Spaziani, M. (2018), "The Statistical Matching of EU-SILC and HBS at ISTAT: Where Do We Stand for the Production of Official

Statistics," Advisory Committee on Statistical Methods, ISTAT. Available at https://www.istat.it/it/files//2018/11/Scanu_original-paper.pdf (accessed August 21, 2022).
European Union (2018), "European Union Statistics for Income and Living Conditions. Cross-Sectional Data." Available at https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions (accessed May 2018).
Eurostat (2013), "Statistical Matching of EU-SILC and the Household Budget Survey to Compare Poverty Estimates using Income, Expenditures and Material Deprivation," Theme: Populations and Social Conditions Collection: Methodologies and Working Papers.
Fosdick, B. K., DeYoreo, M., and Reiter, J. P. (2016), "Categorical Data Fusion Using Auxiliary Information," *The Annals of Applied Statistics*, 10, 1907–1929.
Hu, L., and Bentler, P. M. (1999), "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives," *Structural Equation Modeling*, 6, 1–55.
Kadane, J. B. (2001), "Some Statistical Problems in Merging Data Files," *Journal of Official Statistics*, 17, 423–433.
Kaiser, J. (1983), "The Effectiveness of Hot-Deck Procedures in Small Samples," in Proceedings of the Section on Survey Research Methods, American Statistical Association, Toronto, Canada, pp. 523–528. Available at http://www.asasrms.org/Proceedings/papers/1983_099.pdf.
Lee, K. J., and Carlin, J. B. (2017), "Multiple Imputation in the Presence of Non-Normal Data," *Statistics in Medicine*, 36, 606–617.
Linskens, S. J. (2015), "Statistical Matching: A Comparison of Random and Distance Hot Deck," Report, Tilburg University, The Netherlands.
McMillan, S. (2013), "Comparison of Imputation Methods in the Survey of Income and Program Participation," in Proceedings of the JSM Survey Research Methods Section, Montreal Canada. Available at http://www.asasrms.org/Proceedings/y2013/files/309484_82763.pdf.
Moretti, A., Shlomo, N., and Sakshaug, J. W. (2020), "Multivariate Small Area Estimation of Multidimensional Latent Economic Well-Being Indicators," *International Statistical Review*, 88, 1–28.
Moriarity, C. (2009), *Statistical Properties of Statistical Matching*, Saarbrücken, Germany: VDM Verlag.
Moriarity, C., and Scheuren, F. (2001), "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure," *Journal of Official Statistics*, 17, 407–422.
Morikawa, K., and Kim, J. K. (2018), "A Note on the Equivalence of Two Semiparametric Estimation Methods for Nonignorable Nonresponse," *Statistics & Probability Letters*, 140, 1–6.
Myers, T. A. (2011), "Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data," *Communication Methods and Measures*, 5, 297–310.
Office for National Statistics, Department for Environment, Food and Rural Affairs (2020), *Living Costs and Food Survey, 2017-2018* (3rd ed.), Newport, Wales: UK Data Service. SN: 8459.
Office for National Statistics (2021), "Household Reference Person Harmonised Standard – GSS." https://gss.civilservice.gov.uk/policy-store/household-reference-person/.
Pannekoek, J., Shlomo, N., and De Waal, T. (2013), "Calibrated Imputation of Numerical Data under Linear Edit Restrictions," *Annals of Applied Statistics*, 7, 1983–2006.
Rässler, S. (2004), "Data Fusion: Identification Problems, Validity, and Multiple Imputation," *Austrian Journal of Statistics*, 33, 153–171.
Renssen, R. H. (1998), "Use of Statistical Matching Techniques in Calibration Estimation," *Survey Methodology*, 24, 171–183.
Riddles, M. K., Kim, J. K., and Im, J. (2016), "A Propensity-Score-Adjustment Method for Nonignorable Nonresponse," *Journal of Survey Statistics and Methodology*, 4, 215–245.
Rodgers, W. L. (1984), "An Evaluation of Statistical Matching," *Journal of Business & Economic Statistics*, 2, 91–102.
Rubin, D. B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation," *Journal of Business & Economic Statistics*, 4, 87–94.
Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Shlomo, N., De Waal, T., and Pannekoek, J. (2009), "Mass Imputation for Building a Numerical Statistical Database," paper presented at the UNECE Statistical Data Editing Workshop, Neuchatel, October 2009.

Siddique, J., and Belin, T. R. (2008), "Multiple Imputation Using an Iterative Hot-Deck with Distance-Based Donor Selection," *Statistics in Medicine*, 27, 83–102.

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

Singh, A. C., Mantel, H., Kinack, M., and Rowe, G. (1993), "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption," *Survey Methodology*, 19, 59–79.

Takabe, I., and Yamashita, S. (2020), "New Statistical Matching Method Using Multinomial Logistic Regression Model," in *Advanced Studies in Classification and Data Science. Studies in Classification, Data Analysis, and Knowledge Organization*, eds. T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, and M. Vichi, Singapore: Springer.

Ucar, B., and Betti, G. (2016), "Longitudinal Statistical Matching: Transferring Consumption Expenditure from HBS to SILC Panel Survey," Department of Economics, 739, University of Siena. Available at: https://econpapers.repec.org/paper/usiwpaper/739.htm (accessed July 2019).

van Buuren, S. (2018), *Flexible Imputation of Missing Data* (2nd ed.), Boca Raton: Chapman and Hall/CRC.

van Delden, A., Du Chatinier, B. J., and Scholtus, S. (2020), "Accuracy in the Application of Statistical Matching Methods for Continuous Variables Using Auxiliary Data," *Journal of Survey Statistics and Methodology*, 8, 990–1017.

van Roij, J. (2015), "Statistical Matching: A Comparison of Distance Hot Deck and Model-Based Estimation," Report, Tilburg University, The Netherlands.

Yang, S., and Kim, J. K. (2020), "Statistical Data Integration in Survey Sampling: A Review," *Japanese Journal of Statistics and Data Science*, 3, 625–650.