Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 599–624, http://dx.doi.org/10.2478/JOS-2019-0025

Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data

Eva Endres[1], *Paul Fink*[1], *and Thomas Augustin*[1]

Statistical matching is the term for the integration of two or more data files that share a partially overlapping set of variables. Its aim is to obtain joint information on variables collected in different surveys based on different observation units. This naturally leads to an identification problem, since there is no observation that contains information on all variables of interest.

We develop the first statistical matching micro approach reflecting the natural uncertainty of statistical matching arising from the identification problem in the context of categorical data. A complete synthetic file is obtained by imprecise imputation, replacing missing entries by *sets* of suitable values. Altogether, we discuss three imprecise imputation strategies and propose ideas for potential refinements.

Additionally, we show how the results of imprecise imputation can be embedded into the theory of finite random sets, providing tight lower and upper bounds for probability statements. The results based on a newly developed simulation design – which is customised to the specific requirements for assessing the quality of a statistical matching procedure for categorical data – corroborate that the narrowness of these bounds is practically relevant and that these bounds almost always cover the true parameters.

Key words: Data fusion; data integration; finite random sets; hot deck imputation; (partial) identification.

1. Introduction

Nowadays, a tremendous amount of data is readily accessible, as generated by researchers, companies, and governments. Thus, instead of collecting new data to answer research questions, it is a more convenient alternative to use already available data sources. However, there is often no single data source that includes all information of interest. Statistical matching (also called data integration or data fusion) furnishes a method with which researchers can integrate data collected in different surveys. For example, it was applied by Serafino and Tonkin (2017) to statistically match the data of the *EU Statistics on Income and Living Conditions* and the *Household Budget Survey*.

Assume that we are interested in three blocks of variables, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, while there are two data files, A and B, available. Data file A contains n_{A} observations of $(\boldsymbol{X}, \boldsymbol{Y})$, and data

[1] Ludwig-Maximilians-University of Munich, Ludwigstrasse 33, 80539 Munich, Germany. Emails: eva.endres@stat.uni-muenchen.de, paul.fink@stat.uni-muenchen.de, and augustin@stat.uni-muenchen.de.

Acknowledgments: We thank the associate editor for valuable remarks. We also want to express our gratitude to all participants of the research seminar of the Foundations of Statistics Group at LMU Munich for valuable discussions on this work. The first author also thanks the LMU Mentoring program for financial support.

file B contains n_B observations of $(\boldsymbol{X}, \boldsymbol{Z})$. The observations in B come from the same population but are disjoint from the observations in A. The aim of statistical matching, namely the gain of joint information about variables not jointly observed, is twofold (e.g., D'Orazio et al. 2006b, 2):

(i) the estimation of the joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ or any of its characteristics (*macro approach*), and/or

(ii) the creation of a synthetic data file with complete observations on $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ (*micro approach*).

As the schematic representation in Figure 1 suggests, statistical matching can be interpreted as a missing data problem. The observations of the *specific variables* $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are missing in a special block-wise pattern in $A \Cup B$, which denotes the union of the two available data files. Following, for example, D'Orazio et al. (2006b, 6), the missingness is induced by the given allocation to a certain data file, and thus the missing data mechanism in the framework of statistical matching can convincingly be assumed to be missing completely at random. However, this absence of joint information on all variables leads to A severe identification problem: the parameters that concern the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are not directly estimable from $A \Cup B$. Throughout the article, we use the term *parameter* to refer to a component of the (joint) probability distribution.

For instance, D'Orazio et al. (2006b) show various ways to remedy the issue of non-identifiability. On the basis of their underlying concepts, these methods can be allocated into three basic groups: Approaches which

(i) assume the conditional independence of the specific variables given the *common variables* $\boldsymbol{X}$, in order to achieve a factorisation of the joint distribution whose components are estimable on $A \Cup B$,

(ii) require auxiliary information in terms of a third file or other external information about parameters concerning the relationship of $\boldsymbol{Y}$ and $\boldsymbol{Z}$,

(iii) refrain from aiming at precise point estimates and account for the uncertainty of the statistical matching problem by estimating a set of plausible parameters, resulting in lower and upper bounds for the parameters concerning the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. These estimates can be interpreted as set-valued point estimates, not to be confused with confidence regions.

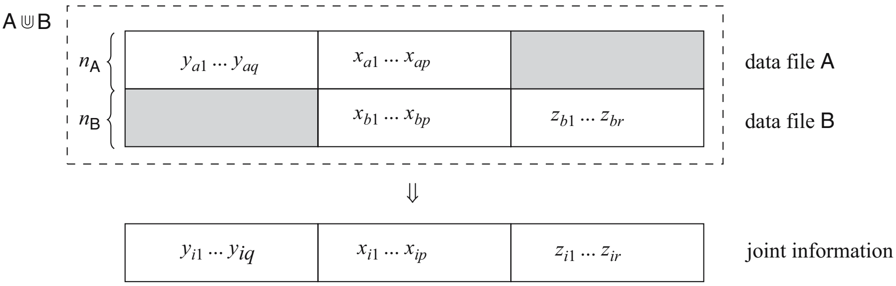

Fig. 1. Schematic representation of the statistical matching problem (See D'Orazio et al. 2006b, 5 (modified)).

In practice, it is not testable whether the conditional independence assumption holds, and in most applications it might be contested. Manski's *Law of Decreasing Credibility* (Manski, 2007, 3), which states that the maintenance of unjustified assumptions reduces the credibility of analyses, makes a very strong argument against the first group of approaches. Auxiliary information, which is the basis of the second group of approaches, is often not available for a certain statistical matching task. Hence, applying statistical matching, taking the underlying uncertainty credibly into account, is the means of choice in these situations.

In the context of statistical matching, typically the term *uncertainty* refers to the previously mentioned identification problem. It points to the fact that even if we have complete information on the marginal distributions of ($\boldsymbol{X}$, $\boldsymbol{Y}$) and ($\boldsymbol{X}$, $\boldsymbol{Z}$), the joint distribution of ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) cannot uniquely be determined (e.g., D'Orazio et al. 2006a). Thus, lower and upper bounds on the parameters (i.e., probability components) are the best that can be obtained without relying on strong untestable assumptions or external information.

Elaborating the concept of uncertainty and how to measure it formed the central focus of the papers by Conti et al. (2012, 2017). Much of the current literature on uncertainty regarding the statistical matching task pays attention to the continuous case, especially to normally distributed variables (e.g., D'Orazio et al. 2006b; Rässler 2002; Ahfock et al. 2016). However, there is also a relatively small body of literature that is concerned with categorical data. For instance, D'Orazio et al. (2006a), Vantaggi (2008), and Di Zio and Vantaggi (2017) deal with statistical matching of categorical data considering different circumstances.

As emphasised by Conti et al. (2012, 70), the "third group of techniques" reflecting the natural uncertainty of statistical matching, does not [usually] "directly aim at reconstructing a complete data set". In the present article, we introduce imprecise (single) imputation as the first micro approach for categorical data that directly accounts for the natural uncertainty of statistical matching. It is based on the imputation of *sets* of plausible values, which leads to a complete synthetic data file with partially set-valued observations. Furthermore, embedding imprecise imputation into the framework of *finite random sets* will allow us to derive lower and upper bounds for the parameters of the joint distribution. As we will highlight, imprecise imputation can be interpreted as a generalisation of multiple hot deck imputation (e.g., Little and Rubin 2002) and fractional hot deck imputation (e.g., Kim and Fuller 2004). The bounds, which we obtain by our imprecise imputation procedure, envelop the results from multiple hot deck imputation and fractional hot deck imputation.

The article is structured as follows. Section 2 recalls the background of our work by giving a brief overview of the basic setting of statistical matching, its interpretation as a missing data problem, and hot deck imputation in this context. Section 3 describes the idea of imprecise imputation and introduces three imputation procedures. Subsequently, in Section 4, we embed imprecise imputation into the theory of finite disjunctive random sets and show how it can be utilised to estimate lower and upper bounds for the parameters of interest from our imputed data file. After providing the setting and results of a simulation study in Section 5, we conclude with a summary and outlook in Section 6. The appendix (Section 7) contains a more detailed description and justification of the design of the simulation study and graphics on the results of the simulation study.

2. Statistical Matching

2.1. *The Basic Setting and its Missing Data Interpretation*

Let us assume that we have two data files, A and B, indexed by $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$, respectively, with n_A and n_B disjoint observation units. Without loss of generality, we assume that the index sets are disjoint: $\mathcal{I}_\mathsf{A} = \{1, \ldots, n_\mathsf{A}\}$ and $\mathcal{I}_\mathsf{B} = \{n_\mathsf{A}+1, \ldots, n_\mathsf{A} + n_\mathsf{B}\}$. Furthermore, let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be the vector of common variables, and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ be the vectors of specific variables. Denote the domains of the possible values of X_l, $l = 1, \ldots, p$, by $\mathcal{X}_l$, their corresponding Cartesian product by $\mathcal{X}$, and proceed analogously for the specific variables, defining $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$, $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$, as well as $\mathcal{Y}$ and $\mathcal{Z}$.

As displayed in Figure 1, data file A exclusively contains information on $(\boldsymbol{X}, \boldsymbol{Y})$ as observations $(\boldsymbol{x}_a, \boldsymbol{y}_a)_{a \in \mathcal{I}_\mathsf{A}}$, while data file B comprises information on $(\boldsymbol{X}, \boldsymbol{Z})$ only, as observations $(\boldsymbol{x}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_\mathsf{B}}$. Consequently, there is no observation that contains simultaneous information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. In the following, the available information will be consolidated in the incomplete sample $\mathsf{A} \Cup \mathsf{B}$, representing the union of files A and B (see Figure 1) with $n := n_\mathsf{A} + n_\mathsf{B}$ observations, indexed by $\mathcal{I} = \mathcal{I}_\mathsf{A} \cup \mathcal{I}_\mathsf{B}$.

Furthermore, we assume that all observations are independently and identically distributed, each following the joint probability distribution $P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z})$, where the realisations for a certain observation $i \in \mathcal{I}$ are depicted as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ir})$. By collecting all probability components of the underlying distribution, we derive the parameter vector consisting of the probability entries of the multidimensional probability table of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

As previously mentioned, statistical matching may be regarded as a missing data problem. Hence, a natural strategy to solve the statistical matching task is imputation, that is, the substitution of the missing entries with suitable real or artificial values to derive a complete (but partially synthetic) data file. To prepare our method, in the following section we focus on *hot deck imputation*, where the missing entries of an observation (*recipient*) are replaced by records from a similar observation (*donor*) of the same sample. Hot deck imputation ensures that only *live* values, that is, actually observed and no artificial values, are substituted, and that the marginal and conditional distributions are preserved well for large samples (e.g., Conti et al. 2008). Hot deck imputation methods are frequently used in practice, comparatively easy to apply, and nonparametric (e.g., Andridge and Little 2010); for a general missing data case, see, for example, Little and Rubin (2002, 66).

2.2. *Hot Deck Imputation for Statistical Matching*

In the context of statistical matching, hot deck imputation belongs to the group of nonparametric micro approaches. In the following, we will recall and formalise an example for four variables (X_1, X_2, Y_1, Z_1) from D'Orazio et al. (2006b, Chap. 2.4) and also explain our notation. The data samples A and B are assigned to the roles of *recipient file* and *donor file*. Since it is a symmetric problem, D'Orazio et al. (2006b) only describe the case where A is the recipient file and B the donor file. The reverse case works analogously. The choice of whether only A, only B, or $\mathsf{A} \uplus \mathsf{B}$ should be imputed depends on many factors. In this article, we impute $\mathsf{A} \uplus \mathsf{B}$ without loss of generality. See, for instance, D'Orazio et al. (2006b, 35–36) for a discussion on this issue.

Random hot deck imputation means that for each missing entry in the recipient file, a donor record from the donor file is randomly chosen by simple random sampling and its corresponding values are used to replace the missing entries in the recipient file. Every missing entry of the specific variable Z_1 in the recipient file A, that is, z_{a1}, $a \in \mathcal{I}_\mathsf{A}$, is replaced by the synthetic value $\tilde{z}_{a1} := z_{b1}$, $b \in \mathcal{I}_\mathsf{B}$, where b is the randomly chosen observation unit from the index set $\mathcal{I}_\mathsf{B}$ of data file B and, hence, $\tilde{z}_{a1} \in \{z_{b1}: b \in \mathcal{I}_\mathsf{B}\}$. The a-th observation of complete, synthetic data file A is composed of $(x_{a1}, x_{a2}, y_{a1}, \tilde{z}_{a1})$, where the tilde marks the imputed and thus synthetic value.

However, simple random sampling gives all observation units in the donor file the same probability of being selected. Thus, it implicitly induces the independence of both the common and specific variables.

A more promising procedure is the assignment of donor and recipient records within groups of similar (homogeneous) records that are created by exploiting the information of the common variables. The realisations of selected categorical common variables are used to generate groups of similar records in both the recipient file and the donor file. Little and Rubin (2002) call these groups *adjustment cells*. Following D'Orazio et al. (2006b), we will call them *donation classes*. The choice of the common variables that are actually used to perform statistical matching (the so-called *matching variables*) has a high impact on the resulting matching quality. It is desirable that the common variables are highly correlated with, or good predictors for the specific variables (Rässler 2002, 10). See, for instance, D'Orazio et al. (2017) on how to choose the *matching variables*.

Consider again data file A as the recipient. The first step of hot deck imputation within homogenous groups is the assignment of all observations in $\mathsf{A} \uplus \mathsf{B}$ to donation classes. For this purpose, we partition the index set $\mathcal{I}$ into $D \leq |\mathcal{X}|$ index sets $\mathcal{I}^d$, $d = 1, \ldots, D$, such that for any d, all observation units in $\mathcal{I}^d$ have the same realisations of $\boldsymbol{X}$. Moreover, define $\mathcal{I}^d_\mathsf{A} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{A}$ and $\mathcal{I}^d_\mathsf{B} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{B}$. Every missing entry for the specific variable Z_1 of an observation unit from A in the d-th donation class, that is, $z_{a1}, a \in \mathcal{I}^d_\mathsf{A}$, is replaced by $\tilde{z}_{a1} := z_{b1},\ \ b \in \mathcal{I}^d_\mathsf{B}$, which is the corresponding value of a randomly chosen observation from the donation class $\mathcal{I}^d_\mathsf{B}$, and hence $\tilde{z}_{a1} \in \left\{z_{b1} : b \in \mathcal{I}^d_\mathsf{B}\right\}$ for all $a \in \mathcal{I}^d_\mathsf{A}$.

Using donation classes, the imputation of Z is conditional on $\boldsymbol{X}$, thus reproducing the empirical conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ in A. Since there are no joint observations of all variables, additionally conditioning on $\boldsymbol{Y}$ is not possible. Thus, a conditional independence – between the imputed values of $\boldsymbol{Z}$ and the values of $\boldsymbol{Y}$, given $\boldsymbol{X}$ – is implicitly (empirically) established in the synthetic parts of the resulting complete file (see Rässler 2002, 200–204).

Every complete synthetic data file that consists of observations $(\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\boldsymbol{z}}_a)_{a \in \mathcal{I}_\mathsf{A}}$ and $(\boldsymbol{x}_b, \tilde{\boldsymbol{y}}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_\mathsf{B}}$ straightforwardly delivers estimates of the underlying joint distribution by evaluating the observed relative frequencies. Written in a form preparing for the generalisation developed in Subsection 4.3, we obtain for an event $\mathcal{E} = \mathcal{E}_\mathcal{X} \times \mathcal{E}_\mathcal{Y} \times \mathcal{E}_\mathcal{Z}$ with $\mathcal{E}_\mathcal{X} \subseteq \mathcal{X}$, $\mathcal{E}_\mathcal{Y} \subseteq \mathcal{Y}$ and $\mathcal{E}_\mathcal{Z} \subseteq \mathcal{Z}$,

$$\begin{aligned}\hat{P}(\mathcal{E}) := \hat{P}((\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \in \mathcal{E}) &= \frac{1}{n}|\{a \in \mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a,\boldsymbol{y}_a,\tilde{z}_a) \in \mathcal{E}\} \cup \{b \in \mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b,\tilde{\boldsymbol{y}}_b,z_b) \in \mathcal{E}\}| \\ &= \frac{1}{n}\big|\{a \in \mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a \in \mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a \in \mathcal{E}_{y}, \tilde{z}_a \in \mathcal{E}_{\mathcal{Z}}\}\big| \\ &\quad + \frac{1}{n}\big|\{b \in \mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b \in \mathcal{E}_{\mathcal{X}}, \tilde{\boldsymbol{y}}_b \in \mathcal{E}_{y}, z_b \in \mathcal{E}_{\mathcal{Z}}\}\big|.\end{aligned} \tag{1}$$

Any event which is not directly representable as a Cartesian product can be decomposed into the union of disjoint events of the previous form.

In the context of missing data, it is a well-known problem that single imputations are not able to reflect the uncertainty that arises from the missingness of joint information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Therefore, it is commonly recommended to apply *multiple imputation* techniques (e.g., Little and Rubin 2002, chap. 5.4), where the replacement of missing entries is performed several times. The obtained complete data files are then analysed by common methods for complete data and the results are subsequently pooled to achieve point estimates. Such multiple imputation techniques have been further developed by Rässler (2002, chap. 4) for application in statistical matching with the intention to estimate lower and upper bounds for the parameters of interest in the spirit of Manski (1995). However, Rässler (2002) only considers normally distributed data and, as stated in Ahfock et al. (2016, 82), by applying multiple imputation "there is no guarantee that the range of imputed datasets fully captures the uncertainty over the partially identified parameters".

3. Imprecise Imputation

3.1. Basic Idea and Terminology

Based on these considerations, we will now develop the concept of imprecise imputation, where we suggest imputing a *set* of plausible values for a missing entry. This leads to precise observations $(\boldsymbol{x}_a, \boldsymbol{y}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\boldsymbol{x}_b, z_b)_{a\in\mathcal{I}_{\mathsf{B}}}$ in B, and to *imprecise*, that is, set-valued, synthetic observations $(\mathfrak{z}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\tilde{\mathfrak{y}}_b)_{b\in\mathcal{I}_{\mathsf{B}}}$in B. Please note that our aim is *not* to identify a single element of these imprecise observations for the purpose of precise single imputation, but rather to regard the whole set as the final piece of indivisible information. In Subsection 4.3 we show how the set-valued imprecise observations can be directly used to obtain estimates for the probability components of the joint distribution.

The following subsections detail and illustrate imprecise imputation. Three different ways of determining the sets of plausible values to be imputed are introduced, each taking into account the variations in how strong and trustworthy the underlying relationship between the common and specific variables is. Without loss of generality, again let A be the recipient and B the donor file, and let the donor classes be defined as in Subsection 2.2.

- **D** *Domain imputation* replaces every missing entry z_{al}, $a \in \mathcal{I}_{\mathsf{A}}$, of a variable Z_l, $l = 1, \ldots, r$, with its domain, that is,

$$\mathfrak{z}_{al} := \mathcal{Z}_l, \quad \forall a \in \mathcal{I}_{\mathsf{A}}, \quad l = 1, \ldots, r. \tag{2}$$

- **VW** *Variable-wise imputation* on the basis of donation classes replaces every missing entry z_{al}, $a \in \mathcal{I}^d_{\mathsf{A}}$, of a variable Z_l, $l = 1, \ldots, r$, with the set of live values of

Z_l within the corresponding class $\mathcal{I}^d_{\mathsf{B}}$. Thus,

$$\tilde{\mathfrak{z}}_{al} := \left\{z_{bl} : b \in \mathcal{I}^d_{\mathsf{B}}\right\}, \quad \forall_a \in \mathcal{I}^d_{\mathsf{A}}, \quad d = 1, \ldots, D, \quad l = 1, \ldots, r. \tag{3}$$

- **CW** *Case-wise imputation*, that is, the simultaneous imputation of all missing entries of an observation a in $\mathcal{I}^d_{\mathsf{A}}$, where every tuple $z_a = (z_{a1}, \ldots, z_{ar})$, $a \in \mathcal{I}^d_{\mathsf{A}}$ is replaced with the set of live tuples in the corresponding class $\mathcal{I}^d_{\mathsf{B}}$. Consequently,

$$\tilde{\mathfrak{z}}_a := \left\{(z_{bl}, \ldots, z_{br}) : b \in \mathcal{I}^d_{\mathsf{B}}\right\}, \quad \forall a \in \mathcal{I}^d_{\mathsf{A}}, \quad d = 1, \ldots, D. \tag{4}$$

3.2. Illustration and Discussion of the Different Types of Imprecise Imputation

3.2.1. Domain Imputation

The most conservative way to determine the set of plausible values that are candidate values for the substitution of a missing entry is to use the whole domain of the corresponding variable. Concretely, this means that every missing entry z_{al}, $a \in \mathcal{I}_{\mathsf{A}}$, $l = 1, \ldots, r$ is substituted by the set of all possible realisations of Z_l, that is, its domain $\mathcal{Z}_l$. Hence, $\tilde{\mathfrak{z}}_{al} := \mathcal{Z}_l, \ \forall a \in \mathcal{I}_{\mathsf{A}}$ becomes a set-valued entry in data file A, where all elements of the set are treated as equally plausible, but without a further reduction in the complexity by some (arbitrary) weighting or aggregation of the elements. The imputed sets for one variable are equal for all observations. This procedure is briefly illustrated in the following running toy example.

Minimal Example 1 *Consider two data files,* A *and* B, *which consist of* $n_{\mathsf{A}} = 2$ *observations of* (Y_1, Y_2, X_1, X_2) *and* $n_{\mathsf{B}} = 3$ *observations of* (X_1, X_2, Z_1, Z_2), *respectively. The corresponding domains of the variables are* $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Z}_1 = \{0,\ 1\}$ *and* $\mathcal{Y}_2 = \mathcal{Z}_2 = \{0,\ 1,\ 2\}$. *Domain imputation results in the following completed data file.*

Table 1. Minimal example 1.

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	{0; 1}	{0; 1; 2}
0	**2**	**0**	**0**	{0; 1}	{0; 1; 2}
{0; 1}	{0; 1; 2}	**1**	**0**	**0**	**0**
{0; 1}	{0; 1; 2}	**1**	**0**	**1**	**1**
{0; 1}	{0; 1; 2}	**0**	**0**	**1**	**2**

Numbers in bold represent the original data. The files A and B are visually divided by the dashed line. The numbers in curly brackets depict the sets of possible realisations of the corresponding variables, that is, the domains, which are here the replacements for the previously missing entries.

This imputation procedure resembles the approach of Ramoni and Sebastiani (2001), who use an incomplete sample to estimate bounds for the parameters of conditional probability distributions in the context of Bayesian networks.

Applying domain imputation, it is guaranteed that the true (but missing) value is always an element of the imputed set. As previously mentioned, domain imputation is very conservative, and thus it can also be applied if the common variables are not good predictors for the specific variables. However, it neglects any available dependence structure between the common and specific variables in the available data. In the following, we will introduce two other methods to determine the set of values for imputation that both take these dependencies into account, albeit to a different extent.

3.2.2. Variable-Wise Imputation

If $q \geq 2$ or $r \geq 2$, with due regard to the association between the common and specific variables, imputation can be performed on two different levels, either by treating each of the specific variables separately or by treating the specific variables within each of the two blocks simultaneously (see, e.g., Joenssen 2015, chap. 3, for precise imputation). In this section, we describe imprecise imputation on the separate level, while the simultaneous level will be addressed in the next section.

The imputation of live values only within donation classes ensures that associations between the common and specific variables are incorporated. As a consequence, the preservation of the dependence structure is improved and the estimated bounds for the parameters of interest become more narrow.

Without loss of generality, again let A be the recipient file and B the donor file. All observations $i \in \mathcal{I}_{\mathsf{A}} \cup \mathcal{I}_{\mathsf{B}}$ are allocated into donation classes depending on their realisations of the matching variables selected from the common variables $\boldsymbol{X}$, following the notation as introduced in Subsection 2.2. For every observation $a \in \mathcal{I}_{\mathsf{A}}^{d}$, the missing entry z_{al} of the variable Z_l, $l = 1, \ldots, r$ is substituted by the set of all live values of this variable from the same donation class in the donor file B, resulting in Equation (3).

Minimal Example 2 *Consider the same data situation as in Example 1. Now we will illustrate the application of the just-described variable-wise imputation. The different backgrounds display the different donation classes based on the combinations of the realisations of X_1 and X_2. Both common variables are used as matching variables in this example.*

Table 2. Minimal example 2.

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	{0; 1}	{0; 1}
0	**2**	**0**	**0**	{1}	{2}
{1}	{2}	**1**	**0**	**0**	**0**
{1}	{2}	**1**	**0**	**1**	**1**
{0}	{2}	**0**	**0**	**1**	**2**

This procedure preserves the dependencies between the common and the specific variables; however, the successive imputation of single variables breaks the dependence structure among the specific variables. Little and Rubin (2002, 72), for instance, have already stated that imputation should be multivariate to preserve the dependencies between the variables. If one attaches high value to this requirement, the imputation should be performed simultaneously for all variables in the data file as described in the following section. Nevertheless, variable-wise imputation is a good compromise between the very conservative domain imputation and the more data-driven case-wise imputation procedure detailed in the following section.

3.2.3. Case-Wise Imputation

For case-wise imputation, we interpret the missing entries of one observation $a \in \mathcal{I}_{\mathsf{A}}^{d}$ out of the d-th donation class in the recipient file as tuple of the form $(z_{a1}, \ldots, z_{ar})$. This tuple of missing entries is replaced by the set of tuples $\mathfrak{z}_a$, which have been observed in the donor file B and the same donation class d, as in Equation (4). This strategy ensures that also the dependencies among the specific variables $\boldsymbol{Z}$ remain unchanged. The following example illustrates this imputation procedure.

Minimal Example 3 *Consider again the situation of Example 1 as a starting point. Interpret the empty cells* z_{a1} *and* z_{a2} *as tuples* (z_{a1}, z_{a2}), $a = 1, 2$, *and analogously* y_{b1} *and* y_{b2} *as tuples* (y_{b1}, y_{b2}), $b = 3, 4, 5$. *The result of case-wise imputation in this example is displayed in the following.*

Table 3. Minimal example 3.

(Y_1, Y_2)	X_1	X_2	(Z_1, Z_2)
(1, 2)	**1**	**0**	{(0, 0); (1, 1)}
(0, 2)	**0**	**0**	{(1, 2)}
{(1, 2)}	**1**	**0**	**(0, 0)**
{(1, 2)}	**1**	**0**	**(1, 1)**
{(0, 2)}	**0**	**0**	**(1, 2)**

3.2.4. General Remarks

A potential issue arises if at least one donation class in the donor file is empty. If so, variable-wise and case-wise imputation cannot directly be applied and we then recommend imputing the domains $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$ or the Cartesian product of the domains $\mathcal{Z}$.

The partially set-valued data files produced by imprecise imputation can be interpreted as a set of underlying precise data files. On closer inspection, the sets produced by the three imputation procedures are nested: the largest set of underlying precise data files is obtained by domain imputation, while case-wise imputation yields the smallest set. Equation (15) shows this relationship formally.

Fractional hot deck imputation (e.g., Kim and Fuller 2004), which is also an imputation approach that is based on set-valued imputations, produces precise results that are contained in the sets obtained by imprecise imputation. It uses a weighting scheme, which is transferred onto the set of values to impute. This strategy reduces complexity by circumventing the direct handling of the imputed set-valued observation by creating a single completed data file with accordingly down-weighted precise pseudo-observations. This kind of precise data allows the direct use of common statistical models and methods. The variability, introduced by having multiple values to be imputed, is accounted for, in the situation of the fractional hot deck imputation, in the variance estimation of the precise estimator. However, variance estimation in the context of fractional hot deck imputation may be argued to be more complex yet more reliable in comparison to multiple imputation (e.g., Yang and Kim 2016).

During the imprecise imputation process, variable-wise and, in particular, domain imputation may create combinations of variable realisations which are contextually unjustified. For instance. D'Orazio et al. (2006b) distinguishes between two types of *logical constraints* to exclude impossible or unlikely combinations in the synthetic categorical data:

(i) *existence of some quantities* on the basis of the individual observation unit, and
(ii) *inequality constraints* on the level of the estimated probability distributions.

Especially the first case can easily be incorporated into the imputation step. Single, implausible values or tuples of values containing the unjustified combinations can easily be removed from the synthetic file. As an extension to both types of constraints, the set of values to be imputed can be restricted further removing not only contextually impossible values but also combinations of values that showed to be very rare within the data file or the population, motivated by the approach of Cattaneo (2013), developed in a decision-theoretic context. This means that the set of (variable-wise or case-wise) live values is restricted to the set of all values whose relative frequencies exceed a certain threshold δ, which may be dependent on the donation class. Increasing δ would gradually eclipse our conservative perspective, resulting, in the extreme case, in a precise single-valued imputation.

We propose to build upon the set-valued data directly, without reducing their complexity via a weighting scheme. In contrast to widely adopted imputation procedures yielding single-valued data, we are now in the situation of statistical analysis of partially set-valued data. To frame imprecise imputation formally, it will be embedded into the concept of finite disjunctive random sets, which allows the estimation of tight lower and upper bounds for the parameters.

In order to allow for a concise description in the following sections, we will take the observation-wise perspective on the imputed sets (i.e., the notation in terms of tuples), which corresponds to the perspective taken by the case-wise imputation. The imputation results of the other procedures can be transferred by taking the Cartesian product, e.g., $\tilde{\boldsymbol{\mathfrak{z}}}_a = \tilde{\mathfrak{z}}_{a1} \times \ldots \times \tilde{\mathfrak{z}}_{ar}$.

4. Imprecision Imputation and Finite Disjunctive Random Sets

Imprecise imputation provides us with partially set-valued data. To prepare a well-founded statistical analysis, we have to formalise imprecise imputation probabilistically. For this purpose, the direct formalisation of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ as collections of random variables and corresponding realisations is no longer sufficient. Starting from an applied point of view, two types of generalisations, which will indeed prove compatible among each other, could be imagined. Firstly, we could abstractly look for a concept of set-valued variables with corresponding set-valued realisations. Secondly, we could assume that every set represents outcomes of various random variables, one of which is the true underlying, yet not precisely observable, random variable. (Throughout this article, we use the term *random variable* to refer to a mapping to the real numbers as well as to some nonnumerical finite space. In the context of the latter, the term *random element* is sometimes used for the sake of distinction e.g., Nguyen 2006).

In this section it will be shown how set-valued observations, and thus the resulting data files of the three imprecise imputation procedures in particular, are covered by the concept of *disjunctive random sets*, also known as *ill-perceived random variables* (Couso et al. 2014; Nguyen 2006). This embedding allows for the assessment of probability statements and the construction of corresponding estimates from the partially set-valued synthetic file derived from imprecise imputation. The interpretation of the set-valued quantities as disjunctive random sets corresponds to the view of Dempster (1967), on which the Dempster-Shafer theory of belief functions (Shafer 1976) is built, which has become very popular in artificial intelligence (see, for example Denoeux 2016).

4.1. Random Set Formulation of Imprecise Imputation

The true random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ map from the underlying population space, denoted by Ω in the sequel, into the domains $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, yielding realisations $\boldsymbol{x}_i$, $\boldsymbol{y}_i$, $\boldsymbol{z}_i$ with $i \in \mathcal{I}$, respectively. Now, neither $\boldsymbol{y}_b$ nor $\boldsymbol{z}_a$ are available, but are replaced by synthetic observations $\tilde{\mathfrak{y}}_b$ and $\tilde{\mathfrak{z}}_a$, respectively, according to either Equation (2), (3), or (4), depending on the chosen imprecise imputation procedure. To formalise this situation, we follow the common practice in statistical matching, treating $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$ as fixed. This allows us to globally replace $\boldsymbol{Y}$ and $\boldsymbol{Z}$ by the set-valued variables $\mathfrak{Y}$ and $\mathfrak{Z}$ (with realisations $\mathfrak{y}_i$ and $\mathfrak{z}_i$, $i \in \mathcal{I}$). The imputed values are already sets, so they fit in nicely, but in order to deal with the already observed realisations, we regard them now as singletons containing only the observed value, for example $\mathfrak{z}_{bl} = \{z_{bl}\}$, $\forall b \in \mathcal{I}_\mathsf{B}, l = 1, \ldots, r$. The variables $\mathfrak{Y}$ and $\mathfrak{Z}$ map into the corresponding power sets $2^{\mathcal{Y}}$ and $2^{\mathcal{Z}}$, whereby mapping into the empty set is excluded.

If we collect the random variables of interest in a variable $\boldsymbol{\Gamma}$ and define $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then

$$\boldsymbol{\Gamma} := (\boldsymbol{X}, \mathfrak{Y}, \mathfrak{Z}) : \Omega \rightarrow 2^{\mathcal{W}} \setminus \{\emptyset\} \tag{5}$$

is a finite nonempty random set (see Definition 3.1 in Nguyen 2006, 35), satisfying the required measurability condition by equipping $2^{\mathcal{W}} \setminus \{\emptyset\}$ with its power set. Since in our setting the imputed (synthetic) set-valued entries of the specific variables are understood as the collection of possible underlying true values, this random set has to be interpreted in the disjunctive way (see, for example Couso et al. 2014; Couso and Dubois 2014).

In general, any disjunctive random set $\boldsymbol{\Gamma}$ induces an upper inverse $\boldsymbol{\Gamma}^*$ and a lower inverse $\boldsymbol{\Gamma}_*$. When considering an event of interest $\mathcal{E} \subseteq \mathcal{W}$, which is now a singleton in the considered space $2^{\mathcal{W}}$, the upper inverse contains all the elements of the population whose image overlaps with $\mathcal{E}$, while the lower inverse contains only those elements of the population whose (nonempty) image is entirely contained within $\mathcal{E}$:

$$\boldsymbol{\Gamma}^*(\mathcal{E}) := \{\omega \in \Omega : \boldsymbol{\Gamma}(\omega) \cap \mathcal{E} \neq \emptyset\} \tag{6}$$

and

$$\boldsymbol{\Gamma}_*(\mathcal{E}) := \{\omega \in \Omega : \boldsymbol{\Gamma}(\omega) \subseteq \mathcal{E}\}. \tag{7}$$

In a heuristic formulation, the upper inverse considers all aspects that do not entirely contradict $\mathcal{E}$, while the lower inverse collects all aspects that necessarily imply $\mathcal{E}$. By using the probability measure $\mathbb{P}$ defined on the original probability space involving Ω, the upper and lower probabilities are then defined in terms of the upper and lower inverse, respectively:

$$P^*(\mathcal{E}) = \mathbb{P}(\boldsymbol{\Gamma}^*(\mathcal{E})) \quad \text{and} \quad P_*(\mathcal{E}) = \mathbb{P}(\boldsymbol{\Gamma}_*(\mathcal{E})) \quad \forall \mathcal{E} \subseteq \mathcal{W}. \tag{8}$$

In order to improve readability we have not marked the image probability measure induced by the random set $\boldsymbol{\Gamma}$, i.e., $P_{\boldsymbol{\Gamma}} = P$, and we proceed analogously with the corresponding set functions P^* and P_*. If we refer to a different image measure, the random quantity inducing this image measure, will be set as subscript to P. If we look at an underlying, ill-perceived random variable $\boldsymbol{W}_0 : \Omega \longrightarrow \mathcal{W}$, only knowing that the unobserved true value $\boldsymbol{W}_0(\omega)$ lies (with probability one) within the observed set $\boldsymbol{\Gamma}(\omega)$, it can be shown (see, for example Couso et al. 2014) that for every event $\mathcal{E} \subseteq \mathcal{W}$ the upper and lower probabilities induced by the random set enclose the probability of $\boldsymbol{W}_0$:

$$P_*(\mathcal{E}) \leq P_{\mathbf{W}_0}(\mathcal{E}) \leq P^*(\mathcal{E}) \quad \forall \mathcal{E} \subseteq \mathcal{W}.$$

This leads to another way of interpreting a random set, namely as producing a family of compatible, precise probability measures $\mathcal{P}(\boldsymbol{\Gamma})$, which is a subset of the set $\mathcal{P}$ of all probability measures on $(2^{\mathcal{W}}, 2^{2^{\mathcal{W}}})$. Nguyen (1978) showed that if $\mathcal{W}$ is finite, the probability distribution induced by $\boldsymbol{\Gamma}$ corresponds to the basic probability assignment in Dempster-Shafer theory and thus makes the belief function mathematically equivalent to P_*. Consequently, the technical results from that area may be used as well.

In the present special case of finite $\mathcal{W}$, the set $\mathcal{P}(\boldsymbol{\Gamma})$ coincides with the credal set $\mathcal{M}(P^*)$, that is, those precise probability measures that respect the upper and lower bounds defined by P^* and P_* event-wise (see Miranda et al. 2010), which also embeds the situation considered here into the framework of imprecise probabilities (e.g., Walley 1991; Augustin et al. 2014).

In particular, P_* and P^* are lower and upper probabilities that are envelopes of all probability measures P in $\mathcal{M}(P^*)$:

$$P_*(\mathcal{E}) = \inf_{P \in \mathcal{M}(P^*)} P(\mathcal{E}) \quad \text{and} \quad P^*(\mathcal{E}) = \sup_{P \in \mathcal{M}(P^*)} P(\mathcal{E}).$$

Indeed, P^*, P_* and $\mathcal{M}(P^*)$ are three mathematically equivalent formulations that can be transferred into each other. Therefore, from an applied point of view, each of them can be seen as the core result of a probabilistic description of imprecise imputation. For any possibly true probability distribution $P_{\mathbf{W}_0}$, our embedding into random sets provides us with a set $\mathcal{M}(P^*)$ of distributions induced by $P_{\mathbf{W}_0}$ such that $\mathcal{M}(P^*)$ contains $P_{\mathbf{W}_0}$. By construction, this is the smallest set that is deducible from the concrete imputation procedure without adding further assumptions or knowledge. Dually, $P^*(\mathcal{E})$ and $P_*(\mathcal{E})$ are the narrowest bounds, deducible on the probabilities of an $\mathcal{E}$.

4.2. Conditioning Disjunctive Random Sets

The representation via the set $\mathcal{M}(P^*)$ of compatible probability distributions including the embedding into the framework of imprecise probabilities guides the further probabilistic analysis of the partially set-valued data file achieved by imprecise imputation. For instance, if the elements of $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ eventually get associated with real-valued outcomes, then a generalised expectation is logically defined via the infimum and supremum of all compatible traditional expectations based on image measures of elements of $\mathcal{M}(P^{*})$.

A similar procedure suggests itself for conditioning, namely an element-wise application of conditioning for all $P \in \mathcal{M}(P^{*})$, provided $P(\mathcal{C}) > 0$ for a conditioning event $\mathcal{C}$ (see, for example Dubois and Prade (1992) or Fagin and Halpern (1991) for a discussion and a comparison to an alternative). It can be shown (e.g., De Campos et al. (1990), Couso et al. (2014), and Fagin and Halpern (1991)) that this leads to the following closed-form results for the upper conditional probability

$$P^{*}(\mathcal{S}|\mathcal{C}) = \sup_{P\in\mathcal{M}(P^{*})} P(\mathcal{S}|\mathcal{C}) = \frac{P^{*}(\mathcal{S}\cap\mathcal{C})}{P^{*}(\mathcal{S}\cap\mathcal{C}) + P_{*}(\bar{\mathcal{S}}\cap\mathcal{C})} \tag{9}$$

and the lower conditional probability

$$P_{*}(\mathcal{S}|\mathcal{C}) = \inf_{P\in\mathcal{M}(P^{*})} P(\mathcal{S}|\mathcal{C}) = \frac{P_{*}(\mathcal{S}\cap\mathcal{C})}{P_{*}(\mathcal{S}\cap\mathcal{C}) + P^{*}(\bar{\mathcal{S}}\cap\mathcal{C})}, \tag{10}$$

where $\bar{\mathcal{S}}$ denotes the complement of $\mathcal{S}$.

4.3. Parameter Estimation by Means of Disjunctive Random Sets Based on Imprecise Imputation

So far, this approach has been described in a probabilistic setting, where every entity involved is known (besides the true hidden/ill-perceived random variable). In the following, the statistical perspective will be taken in which the probabilities that correspond to the random set need to be estimated from a finite sample. Consequently, we take our synthetic data file derived from imprecise imputation as consisting of $n = n_{\mathsf{A}} + n_{\mathsf{B}}$ realisations γ_i, $i \in \mathcal{I}$, of the corresponding generic random set $\boldsymbol{\Gamma}$ from Equation (5). Referring to Equation (8), with Equations (6) and (7), we obtain, in a generalisation of Equation (1), for our event $\mathcal{E} = \mathcal{E}_{\mathcal{X}}\times\mathcal{E}_{\mathcal{Y}}\times\mathcal{E}_{\mathcal{Z}}$:

$$\begin{aligned}\widehat{P^{*}}(\mathcal{E}) &= \frac{1}{n}\left|\{i\in\mathcal{I} : \gamma_i\cap\mathcal{E}\neq\emptyset\}\right| \\ &= \frac{1}{n}\Big(\left|\{a\in\mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a,\boldsymbol{y}_a,\tilde{\boldsymbol{\mathfrak{z}}}_a)\cap\mathcal{E}\neq\emptyset\}\right| + \left|\{b\in\mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b,\tilde{\boldsymbol{\mathfrak{y}}}_b,\boldsymbol{z}_b)\cap\mathcal{E}\neq\emptyset\}\right|\Big) \\ &= \frac{1}{n}\Big|\{a\in\mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a\in\mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a\in\mathcal{E}_{\mathcal{Y}}, \tilde{\boldsymbol{\mathfrak{z}}}_a\cap\mathcal{E}_{\mathcal{Z}}\neq\emptyset\}\Big| \\ &\quad + \frac{1}{n}\Big|\{b\in\mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b\in\mathcal{E}_{\mathcal{X}}, \tilde{\boldsymbol{\mathfrak{y}}}_b\cap\mathcal{E}_{\mathcal{Y}}\neq\emptyset, z_b\in\mathcal{E}_{\mathcal{Z}}\}\Big|\end{aligned} \tag{11}$$

and

$$\begin{aligned}\widehat{P_*}(\mathcal{E}) &= \frac{1}{n}|\{i \in \mathcal{I} : \gamma_i \subseteq \mathcal{E}, \gamma_i \neq \emptyset\}| \\ &= \frac{1}{n}\left(|\{a \in \mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\mathfrak{z}}_a) \subseteq \mathcal{E}\}| + |\{b \in \mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b, \tilde{\mathfrak{y}}_b, \boldsymbol{z}_b) \subseteq \mathcal{E}\}|\right) \\ &= \frac{1}{n}\left|\{a \in \mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a \in \mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a \in \mathcal{E}_{\mathcal{Y}}, \tilde{\mathfrak{z}}_a \subseteq \mathcal{E}_{\mathcal{Z}}\}\right| \\ &\quad + \frac{1}{n}\left|\{b \in \mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b \in \mathcal{E}_{\mathcal{X}}, \tilde{\mathfrak{y}}_b \subseteq \mathcal{E}_{\mathcal{Y}}, z_b \in \mathcal{E}_{\mathcal{Z}}\}\right|.\end{aligned} \tag{12}$$

From $\widehat{P^*}(\mathcal{E})$ and $\widehat{P_*}(\mathcal{E})$ also an estimate of the induced underlying set of probability measures can be derived:

$$\widehat{\mathcal{M}}(P^*) = \{P \in \mathcal{P} : \widehat{P_*}(\mathcal{E}) \leq P(\mathcal{E}) \leq \widehat{P^*}(\mathcal{E}), \quad \forall \mathcal{E} \subseteq \mathcal{W}\}. \tag{13}$$

In comparing the estimates resulting from the different types of imputation procedures, it is essential to recall that the different set-valued data files are nested, by construction, with respect to all compatible underlying precise data files. The set resulting from domain imputation is a (nonstrict) superset of the set obtained from variable-wise imprecise imputation, which contains the set produced by case-wise imprecise imputation. Therefore, with the abbreviations introduced in Subsection 3.1, it holds that

$$\widehat{\mathcal{M}}\left(P^{*^{CW}}\right) \subseteq \widehat{\mathcal{M}}\left(P^{*^{VW}}\right) \subseteq \widehat{\mathcal{M}}\left(P^{*^{D}}\right) \tag{14}$$

and, for every event $\mathcal{E} \subseteq \mathcal{W}$,

$$\widehat{P_*}^{D}(\mathcal{E}) \leq \widehat{P_*}^{VW}(\mathcal{E}) \leq \widehat{P_*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{VW}(\mathcal{E}) \leq \widehat{P^*}^{D}(\mathcal{E}). \tag{15}$$

This allows us to compare the results obtained through the different imputation approaches to the result under conditional independence, which yields a single precise probability distribution. It can be argued that the probability distribution under conditional independence is contained in any of the estimated sets. Furthermore, as can be seen from the relations between the different sets of probabilities in Equation (14), the set induced by case-wise imputation can be regarded as containing probability distributions neighbouring the one under conditional independence. The other sets can be interpreted to deviate even more from conditional independence, where domain imputation has the largest deviation. Domain imputation demonstrably neglects any conditional dependence structure in the construction of its bounds. Therefore, the bounds are maximal, but not vacuous, thus constraining the parameter space.

In addition to logical constraints on the imputation level (see Subsubsection 3.2.4), constraints on the level of the estimated probability distribution can be regarded as a refinement of the estimated set $\widehat{\mathcal{M}}(P^*)$ of probabilities derived from our imprecise imputation (see Equation (13)). Since by construction $\widehat{\mathcal{M}}(P^*)$ is representable as a convex polyhedron in $\mathbb{R}^{|\mathcal{W}|-1}$, especially linear constraints can be incorporated very conveniently.

Minimal Example 4 *For demonstration purposes, let us estimate the bounds of conditional probabilities $P(Y_1 = 1|Z_1 = 1)$ for the case-wise imputed data of our toy* *example from Example 3. For the upper conditional probability we need to estimate* $P^{*}(Y_1 = 1, Z_1 = 1)$ *and* $P_{*}(Y_1 \neq 1, Z_1 = 1)$ *in accordance to Equation* (9). *We estimate the upper joint probability with Equation* (11) *by counting how many observations have or could have realisation with* $y_1 = 1$ *and* $z_1 = 1$. *This holds for observations 1 and 4:* $\widehat{P^{*}}(Y_1 = 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$. *The lower joint probability is obtained by Equation* (12) *by counting how many observations only have realisations with* $Y_1 \neq 1$ *and* $Z_1 = 1$. *This holds for observations 2 and 5, and hence* $\widehat{P_{*}}(Y_1 \neq 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$ *and thus the upper conditional probability is* $\widehat{P^{*}}(Y_1 = 1 | Z_1 = 1) = \frac{0.4}{0.4+0.4} = 0.5$. *Similarly, the lower and upper joint probabilities are estimated, occurring in Equation* (10): $\widehat{P_{*}}(Y_1 = 1, Z_1 = 1) = 0.2$ and $\widehat{P^{*}}(Y_1 \neq 1, Z_1 = 1) = 0.4$, *resulting in the lower conditional probability* $\widehat{P_{*}}(Y_1 = 1 | Z_1 = 1) = \frac{0.2}{0.4+0.2} = \frac{1}{3}$. *Thus,* $\hat{P}(Y_1 = 1 | Z_1 = 1)$ *is within the interval* $\left[\frac{1}{3}; \frac{1}{2}\right]$.

5. Simulation Study of Imprecise Imputation

To investigate the quality of imprecise imputation, we have performed a simulation study. It would have been possible to also match real data, but in a real-data application the true underlying distribution is unknown and assessing the statistical matching quality is possible only by checking whether the marginal distributions are preserved. Since this is clearly not sufficient as a sole quality criterion, we have simulated data. With the aid of a simulation study we have also been able to cover various data scenarios which make the results of our investigation of the quality criteria more credible. Moreover, the noise arising from the sampling procedure in the context of real-data applications is neutralised.

We simulated a complete categorical data file $\mathsf{A} \uplus \mathsf{B}$ with i.i.d. observations and split it into two separate files, A and B, with $n_{\mathsf{A}} = n_{\mathsf{B}}$. Subsequently, the observations of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are deleted from A and B, respectively, and the two files are statistically matched by imprecise imputation. To assess the statistical matching quality, we analysed, on the one hand, whether the true parameters of the marginal distributions and the joint distributions are within their respective estimated bounds, and, on the other hand, the distance between the upper and lower bounds. This distance, which we will call *interval width* in the following, is an appropriate performance measure since the true parameters would always lie within the estimated bounds if we chose the unit interval as a trivial estimator of a probability component. Thus, the narrower the interval that covers the component of the true parameter, the better the procedure performs. In the following, we will detail the simulation design, parameters, and results. All simulations and analyses are conducted in R (R Core Team 2018). The specific task presented in this paper is implemented in a published R-package *impimp* (Fink et al. 2019), which was also utilised in the simulation, but is in the same way usable for real-data applications.

5.1. Simulation Design

The starting point of our simulation analysis is two categorical data files, A and B. Both of them contain information on four common variables $\boldsymbol{X} = \{X_1, X_2, X_3, X_4\}$ and four specific variables $\boldsymbol{Y} = \{Y_1, Y_2, Y_3, Y_4\}$ or $\boldsymbol{Z} = \{Z_1, Z_2, Z_3, Z_4\}$, respectively, with domains $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Z}_1 = \mathcal{Z}_2 = \{0, 1\}$ and $\mathcal{X}_3 = \mathcal{X}_4 = \mathcal{Y}_3 = \mathcal{Y}_4 = \mathcal{Z}_3 = \mathcal{Z}_4 = \{0, 1, 2\}$.

Altogether, we modify the following four simulation parameters:

1. The strength of the bivariate associations in terms of the corrected contingency coefficient C, also known as Sakoda's adjusted Pearson's C: $C \in [0, 0.2)$, $C \in [0.2, 0.6)$, or $C \in [0.6, 1)$;
2. The Jensen-Shannon divergence *JSD* (e.g., Lin 1991) from the marginal distribution of the common variables to the discrete uniform distribution: $JSD > 0.15$ or $JSD \leq 0.015$;
3. The numbers of observations $n_{\mathsf{A}} = n_{\mathsf{B}} \in \{50, 100, 250\}$; and
4. the dependence structure among the variables (see Figure 2).

Altogether, we obtain 72 simulation scenarios. An explanation of the choice of the simulation parameters follows in the next section. An exhaustive justification and description of the simulation design can be found in Appendix A and Appendix B, respectively.

5.2. Simulation Parameters

As already stated by Rässler (2002, 10), the common variables should be good predictors for the specific variables. This ensures that the donation classes are suitable for generating homogeneous groups of observations that lead to proper donor values for a missing entry. Taking this fact into account, we vary the dependence structure within a simulated data file in terms of its bivariate associations.

Figure 2 shows four different dependence structures that are covered by our simulation design. The upper six variables of each design represent the binary variables, and the six variables below the dashed line represent the variables with three categories. The connecting lines between the variables display the bivariate dependencies among these variables. For example, in the top line of Structure 1, the variable X_1 is connected to variable Y_1 and also to variable Z_1. The strengths of these bivariate associations are controlled by the corrected contingency coefficient $C \in [0, 1]$. This association measure for categorical variables is based on the $\mathcal{X}^2$-coefficient for contingency tables, but is corrected for the number of observations, as well as for the number of categories.

At first sight, the number of observations plays a counterintuitive role in this simulation study. We expect that the distances between the lower and upper bounds of the parameters of interest increase in situations with a higher number of observations. This is due to the

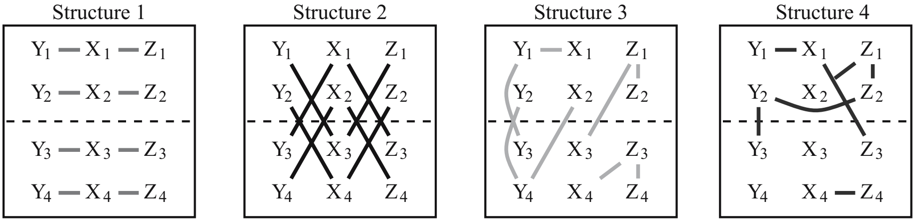

Fig. 2. Four different dependence structures among the variables in the simulation study. A line between two variables indicates dependence between them.

fact that a growth of the number of observations also causes an increase in the number of missing entries, which, in turn, leads to less precise estimations.

The Jensen-Shannon divergence from the marginal distributions of the common variables to the discrete uniform distribution is expected to have an indirect effect on the statistical matching quality. If one or more of these marginals are far away from the discrete uniform distribution, we obtain rare realisations of our matching variables, which induce rare donation classes. This circumstance may likely lead to situations where certain rare donation classes of the recipient file do not exist in the donor file. In these cases, we impute, in accordance with the recommendation in Subsubsection 3.2.4, the domain for the missing entries that corresponds to a minimum of information which, in turn, leads to bounds that are (slightly) further apart.

5.3. Simulation Results

As discussed, we use two measures of quality. Firstly, we investigate whether the true parameters of our simulation distributions lie within the corresponding lower and upper bounds estimated on the synthetic and partially set-valued data. Secondly, we report the mean interval widths that equal the mean distances between the upper and lower bounds. An interval width of 0 corresponds to a precise estimation.

Table 4 shows that the true values of the components of the marginal and the joint distributions almost always lie inside the estimated bounds. When considering the coverage of the marginal distributions (upper part of Table 4), the only visible difference is between the domain and donation-based approaches with respect to the coverage of the true probability: while the intervals for domain imputation are always wide enough to cover the true probability, for variable-wise and to the same extent for case-wise imputation the estimated intervals are sometimes too narrow. Regarding the joint distribution (lower part of Table 4), the intervals estimated on the domain-imputed data still always cover the true probability, but there is now also a slight difference between

Table 4. Relative number of probability table components for which the true parameter of the marginal distributions (top) / joint distributions (bottom) lies inside the estimated bounds, aggregated over all repetitions. The presented summary lists the result when pooling all simulation scenarios. The absence of decimal places for domain imputation highlights the numerically exact values.

Imputation procedure	Min.	1st quartile	Median	3rd quartile	Max.	Mean
Domain	1	1	1	1	1	1
Variable-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792
Case-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792

Imputation procedure	Min.	1st quartile	Median	3rd quartile	Max.	Mean
Domain	1	1	1	1	1	1
Variable-wise	0.9975	0.9989	0.9994	0.9996	0.9998	0.9992
Case-wise	0.9944	0.9985	0.9990	0.9993	0.9997	0.9987

case-wise and variable-wise imputation, showing the hierarchy of the intervals as given in Equation (15). Nonetheless, the estimated intervals of the donation-based imputation approaches still almost always cover the true probability. The difference between marginal and joint coverage is mostly due to the fact that, using the simulation design, the joint distribution had more components (46,656) than observations in the data file, which means that most of the underlying probability entries were zero. The marginal distributions, in contrast, consisted of only two to three entries, which made it harder to distinguish on the estimated level between the different imputation approaches. By and large, the results show a desirable output and also demonstrate the power of our method, which achieves high average coverage even across the diverse simulation scenarios.

The interval width was separately analysed for the components of the marginal distributions and joint distributions within the simulation. The aggregated results are displayed in the figures in Appendix C and summarised in the following.

The mean and maximal interval widths of the estimated intervals for the marginal distributions using domain imputation are always 0.5. This is the maximum interval width which can be achieved if we impute $\mathsf{A} \Cup \mathsf{B}$ under the constraint that $n_{\mathsf{A}} = n_{\mathsf{B}}$. Both variable-wise imputation and case-wise imputation yield intervals that are, in most of the cases, smaller than the intervals obtained by domain imputation. This also holds for the components of the joint distributions.

The interval widths of the marginals are conspicuously affected by the divergence of the marginal distributions to the discrete uniform distribution. If the marginals are close to the uniform distribution, the intervals are narrow. However, this effect decreases if there are few direct connections between the specific variables and the common variables. For the interval widths of the components of the joint distribution, we can observe a slightly contrary effect regarding the combination of marginals that are close to the uniform distribution and few direct connections between the specific variables and common variables. For the simulation designs with a higher divergence to the uniform distribution, the variation of the interval widths is considerably smaller. Moreover, in these cases, the median of the interval widths lies below the median of the design, with a smaller divergence to the uniform distribution. At first sight, this result appears somewhat counterintuitive, but can be explained as follows. Given a fixed value for the corrected contingency coefficient C, with marginal distributions of the common variables which are far away from the discrete uniform distribution, we obtain a probability table which has fewer combinatorial possibilities for each cell than with marginals close to the uniform distribution. This circumstance makes the estimation more precise in some cases, which in turn leads to smaller interval widths.

Furthermore, the results show that with a growing number of observations, the interval widths of the marginal distributions slightly increase. The interval widths also show higher variations in these cases. The interval widths for the components of the joint distribution show the same behaviour with respect to the number of observations.

The strengths of the bivariate associations in terms of the corrected contingency table also affect the widths of the intervals concerning the marginal distributions. In particular, the first dependence structure shows that the interval width decreases with a higher C. Nevertheless, the difference between low and high associations is, in few cases, (especially for marginals close to the uniform distribution) opposite, or only visible in the variations. Considering the interval widths for the components of the joint distribution, we can see that high associations improve the estimation.

The simulation results also show that, as expected, the dependence structure among the variables in a data file has an influence on the estimated lower and upper bounds of the parameters of the marginal distributions. The mean interval widths increase if the specific variables and the common variables have only few connections. The last dependence structure where there are only few connections between the common variables and the specific variables tends to lead to intervals with higher widths for the components of the joint distribution.

To sum up, all imputation procedures yield lower and upper bounds that almost always cover the components of the true parameter value. The number of cases where a component of the true parameter lies outside of the estimated interval is negligible. Additionally, the width of the intervals decreases the more the dependence structure among the variables in the data file are incorporated in the imputation procedure. This also holds for small associations and for structures where the specific variables only have few connections to the common variables.

6. Concluding Remarks

We have presented the first micro approach for statistical matching of categorical data that reflects the natural uncertainty of statistical matching. Our approach relies on imprecise imputation, that is, the idea to impute sets of plausible values. We suggested three types of imputation strategies: domain, variable-wise, and case-wise imprecise imputation. They can be distinguished by their ability to reproduce the available dependence structure between the common and the observed specific variables in the originals files A and B into the synthetic file. They also differ in the amount of data constellations produced beyond those obtained by single or multiple imputation under the conditional independence assumption. Imprecise imputation can be seen as a set-valued generalisation of multiple (hot deck) imputation on the one hand, and fractional hot deck imputation on the other hand.

The most conservative approach, domain imputation, does not take any dependencies in the original data into account. Essentially, the dependencies present in the original files are diluted in the resulting complete synthetic file. This approach is suitable especially when there is little dependence between the common and specific variables. On the other hand, imprecise imputation based on donation classes is able to utilise the observed dependencies between the common and specific variables, and even, in the example of the case-wise variant, within the specific variables.

Embedding imprecise imputation into the framework of finite random sets allows us to derive set-valued estimates of the underlying true parameters. These estimates – possibly after their refinement by external information, see, for example, Subsubsection 3.2.4 – reflect the uncertainty inherent in the identification problem of statistical matching. The estimation procedure utilises the set-valued information to full extent without artificially reducing the complexity of the imputed sets. Simulation results, based on a new simulation technique for dependent categorical data, corroborate that the true parameter values lie almost always inside the respective estimated bounds.

Imprecise imputation is an intuitive statistical matching micro approach which can easily be extended for more than two data files. In a strongly unbalanced statistical matching situation where, for example $n_{\mathsf{A}} \ll n_{\mathsf{B}}$, imprecise imputation can be applied straightforwardly to impute only the smaller file. If so, A takes the role of the recipient and the larger file, B, the role of the donor. In this special situation, the estimates for the specific variables $\boldsymbol{Y}$ are precise.

Moreover, the imprecise imputed data file with synthetic set-valued observations can be used as a starting point to derive one or multiple data files of the usual form. This would bring back the opportunity to use statistical procedures for the analysis of these now entirely single-valued data and to combine the results obtained from those data files by common multiple imputation techniques. However, one would then lose sight, to a considerable extent, of the conviction of this work, which is to produce a credible analysis by taking the full uncertainty into account.

Further studies need to be carried out to validate the performance of imprecise imputation. On the one hand, additional simulation parameters and dependence structures should be investigated in simulation studies. On the other hand, the performance of imprecise imputation should also be assessed by real-data applications. However, considerably more work will need to be done to find a definition of appropriate statistical matching quality criteria, since the true joint distribution is not available for comparisons. A further natural progression of this work is the comparison of imprecise imputation to existing statistical matching macro approaches that also address the identification problem. For this purpose, a comparison of the uncertainty measures introduced in Conti et al. (2012) or Conti et al. (2017) is desirable.

Finally, we should stress that imprecise imputation is not restricted to the block-wise missing pattern in the statistical matching framework: it is also applicable to general missing data problems. All three types of imprecise imputation promise considerable potential for a credible analysis of (non)randomly missing data far beyond statistical matching and are worthwhile to be elaborated upon and evaluated in detail.

7. Appendix

7.1. *Appendix A. Why we need a new simulation procedure*

To generate simulated categorical data that meet all the desired properties, we propose a new procedure which we detail in the following section. However, first we want to elucidate why conventional simulation approaches are not suitable for our requirements. The key aspects are listed as follows:

(i) One way to generate categorical data with predefined properties is to draw random observations from a multidimensional probability table, which, on the one hand, fulfils all of these properties that, on the other hand, represents the probability entries of the joint distribution of all variables. The main disadvantage of this procedure is that it can be very difficult to find a suitable joint distribution that fulfils all the desired properties. Furthermore, we would argue that it is necessary to consider several joint distributions in order to draw valid conclusions about the performance of imprecise imputation, which in turn makes the problem of finding suitable distributions even harder.

(ii) Another option would be the simulation of categorical data based on a multidimensional (logit) regression model. However, a regression model cannot be used to control for the dependence structure and strength within the set of variables in the detail we wish to have.

(iii) The simulation of categorical data which imply a certain dependence structure can also be realised using a probabilistic graphical model such as a Bayesian network. The major problem with this way of proceeding is the resulting conditional independence among parts of our variables. If the – in real-world applications potentially unjustified – conditional independence assumption holds in our simulated data, statistical matching techniques directly utilising this assumption would unfairly outperform, making a fair comparison of procedures impossible.

(iv) A further feasible way to generate dependent categorical data is to employ a multivariate normal distribution with a predefined correlation matrix and discretise the data drawn from it. Nevertheless, the resulting simulated data have an ordinal scale instead of a nominal scale and we have no direct control on the strengths of the dependencies in terms of the corrected contingency coefficient. The same problems hold for simulation techniques that are based on a Gaussian copulas, such as the one suggested by Barbiero and Ferrari (2017).

To sum up, our goal is to use a simulation technique that takes all of our desired properties into account and avoid the problems described previously.

7.2. *Appendix B. Simulation procedure*

For this purpose, we invented a new simulation procedure that is directly based on two-way tables of relative frequencies and a suitable association measure. The bivariate associations within the simulated data can be expressed by this association measure on

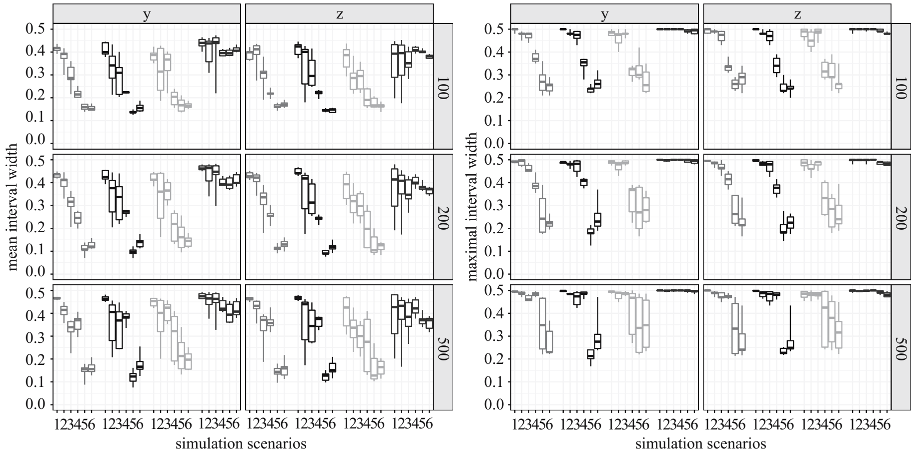

*Fig. 3. Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables **Y** and **Z**, respectively.*

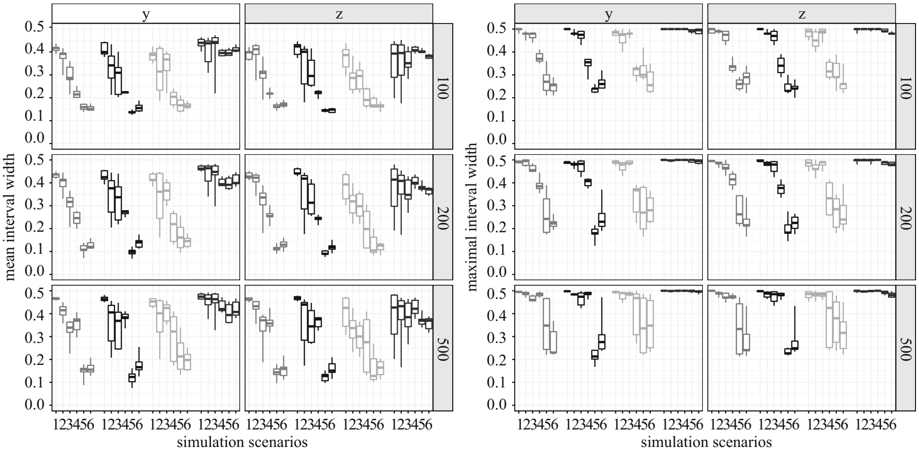

Fig. 4. Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables **Y** *and* **Z**, *respectively.*

bivariate frequency tables of sizes 2 × 2, 2 × 3, and 3 × 3 reflecting the domains listed in Section 5. As also mentioned therein, we use the corrected contingency coefficient to express the strength of associations. Since – for a fixed and known number of observations – the absolute frequencies can be directly derived by the relative frequencies, and vice versa, this association measure is also suitable for tables of relative frequencies and leads to the same results.

In a first step, we generate a set S of relative frequency tables that represents the set of all possible frequency tables of above-mentioned sizes. S is created by taking all combinations of two discrete (marginal) probability distributions, whose event probabilities are strictly positive and on a one-percent grid. This strict positivity is needed because zero entries in the marginal distributions lead to zero entries in the table under independence. This entails that the $\mathcal{X}^2$ coefficient and all association measures based

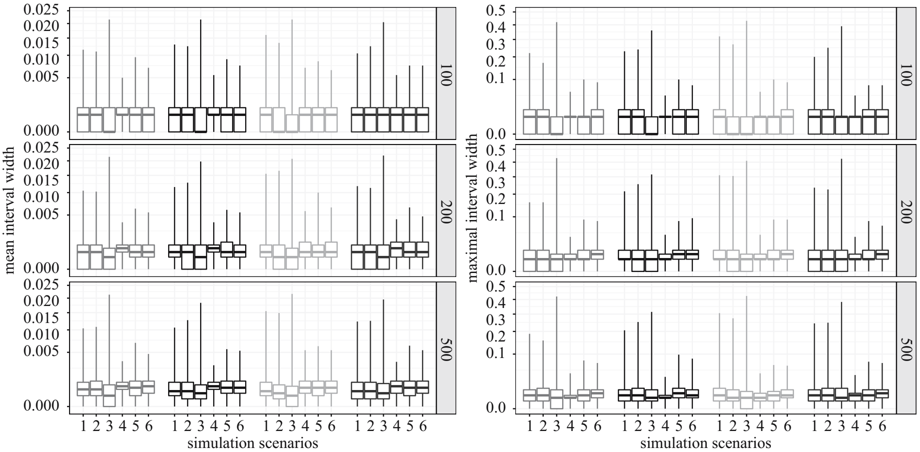

Fig. 5. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of **X**, **Y**, **Z** *for domain imputation.*

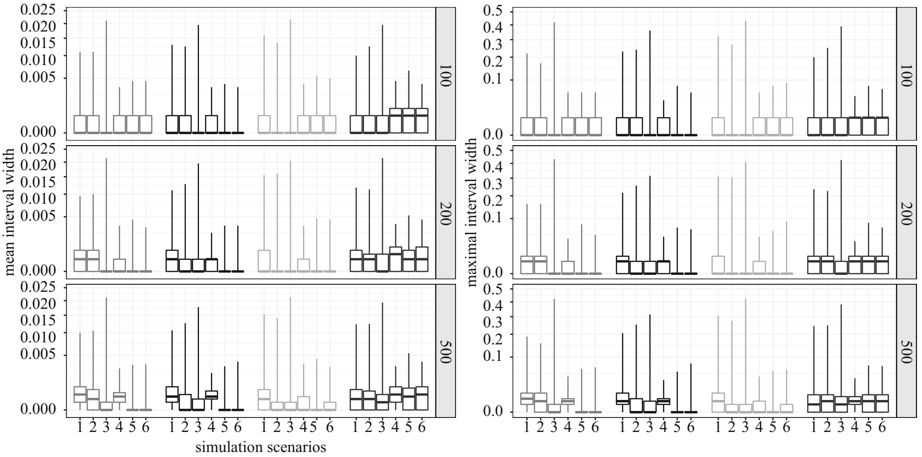

*Fig. 6. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of **X**, **Y**, **Z** for variable-wise imputation.*

on it are not defined. S covers a large variety of marginal distributions and association measures ($|\mathrm{S}| = 48\ 044\ 502$).

In a second step, we randomly draw one frequency table from S^* for each bivariate association depicted in Figure 2, where $S* \subseteq S$ denotes the set of probability tables that meets all predefined requirements for a specific simulation setting. Afterwards, we multiply the selected tables of relative frequencies with the desired number of observations and create a data file with complete observations $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. To meet the challenges of a statistical matching framework, we split this data file into two parts that represent the files A and B with $n_{\mathsf{A}} = n_{\mathsf{B}}$, and remove the observations $\boldsymbol{z}$ from A and $\boldsymbol{y}$ from B, respectively.

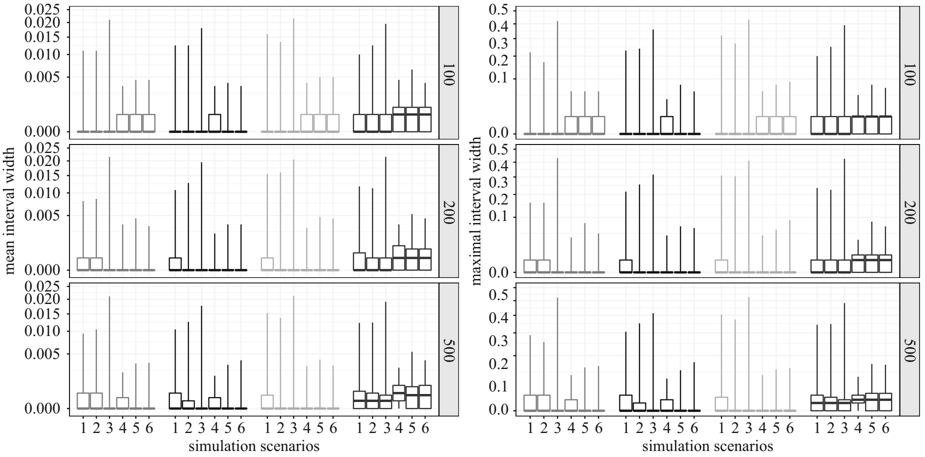

*Fig. 7. Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of **X**, **Y**, **Z** for case-wise imputation.*

7.3. Appendix C. Simulation results

Figures 3–7 show the interval widths of the parameter estimates on the partially set-valued synthetic data, aggregated for 20 simulation runs. The graphics are grouped by the different dependence designs (see Figure 2) and the numbers of observations. The results are displayed separately for the parameters of the marginal distributions and the parameters of the joint distributions. The whiskers range from the minimum to the maximum to ensure better readability. Please note that while the interval widths for the components of the joint distribution are reported on a square root scale to spread the values and make the different results more visible, the values themselves are not transformed.

The figure showing the mean and maximal interval widths of the components of the marginal distributions of the specific variables for domain imputation is not shown here since the interval widths are 0.5 for all simulation scenarios. This is no coincidence and results deterministically from the numbers of observations n_{A} and n_{B}.

8. References

Ahfock, D., S. Pyne, S.X. Lee, and G.J. McLachlan. 2016. “Partial Identification in the Statistical Matching Problem.” *Computational Statistics & Data Analysis* 104: 79–90. Doi: https://doi.org/10.1016/j.csda.2016.06.005.

Andridge, R.R. and R.J.A. Little. 2010. “A Review of Hot Deck Imputation for Survey Nonresponse.” *International Statistical Review* 78: 40–64. Doi: https://doi.org/10.1111/j.1751-5823.2010.00103.x.

Augustin, T., Coolen, F.P.A., de Cooman, G., and Troffaes, M.C.M. (Eds.). 2014. *Introduction to Imprecise Probabilities*. Chichester: Wiley. Doi: https://doi.org/10.1002/9781118763117.

Barbiero, A. and P.A. Ferrari. 2017. “An R Package for the Simulation of Correlated Discrete Variables.” *Communications in Statistics – Simulation and Computation* 46: 5123–5140. Doi: https://doi.org/10.1080/03610918.2016.1146758.

Cattaneo, M. 2013. “Likelihood Decision Functions.” *Electronic Journal of Statistics* 7: 2924–2946. Doi: https://doi.org/10.1214/13-EJS869.

Conti, P.L., D. Marella, and M. Scanu. 2008. “Evaluation of Matching Noise for Imputation Techniques Based on Nonparametric Local Linear Regression Estimators.” *Computational Statistics & Data Analysis* 53: 354–365. Doi: https://doi.org/10.1016/j.csda.2008.07.041.

Conti, P.L., D. Marella, and M. Scanu. 2012. “Uncertainty Analysis in Statistical Matching.” *Journal of Official Statistics* 28: 69–88. Available at: http://www.scb.se/dokumentation/statistiska-metoder/JOS-archive/ (accessed July 2019).

Conti, P.L., D. Marella, and M. Scanu. 2017. “How Far from Identifiability? A Systematic Overview of the Statistical Matching Problem in a Non Parametric Framework.” *Communications in Statistics Theory and Methods* 46: 967–994. Doi: https://doi.org/10.1080/03610926.2015.1010005.

Couso, I. and D. Dubois. 2014. “Statistical Reasoning with Set-valued Information: Ontic vs. Epistemic Views.” *International Journal of Approximate Reasoning* 55: 1502–1518. Doi: https://doi.org/10.1016/j.ijar.2013.07.002.

Couso, I., D. Dubois, and L. Sánchez. 2014. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Cham: Springer. Doi: https://doi.org/10.1007/978-3-319-08611-8.

De Campos, L.M., M.T. Lamata, and S. Moral. 1990. "The Concept of Conditional Fuzzy Measure." *International Journal of Intelligent Systems* 5: 237–246. Doi: https://doi.org/10.1002/int.4550050302.

Dempster, A.P. 1967. "Upper and Lower Probabilities Induced By a Multivalued Mapping." *The Annals of Mathematical Statistics* 38: 325–339. Doi: https://doi.org/10.1214/aoms/1177698950.

Denoeux, T. 2016. "40 Years of Dempster-Shafer Theory." *International Journal of Approximate Reasoning* 79: 1–6. Doi: https://doi.org/10.1016/j.ijar.2016.07.010.

Di Zio, M. and B. Vantaggi. 2017. "Partial Identification in Statistical Matching with Misclassification." *International Journal of Approximate Reasoning* 82: 227–241. Doi: https://doi.org/10.1016/j.ijar.2016.12.015.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006a. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints." *Journal of Official Statistics* 22: 137–157. Available at: http://www.scb.se/dokumentation/statistiska-metoder/JOS-archive/ (accessed July 2019).

D'Orazio, M., M. Di Zio, and M. Scanu. 2006b. *Statistical Matching: Theory and Practice*. Chichester: Wiley. Doi: https://doi.org/10.1002/0470023554.

D'Orazio, M., M. Di Zio, and M. Scanu. 2017. "The Use of Uncertainty to Choose Matching Variables in Statistical Matching." *International Journal of Approximate Reasoning* 90: 433–440. Doi: https://doi.org/10.1016/j.ijar.2017.08.015.

Dubois, D. and H. Prade. 1992. "Evidence, Knowledge, and Belief Functions." *International Journal of Approximate Reasoning* 6: 295–319. Doi: https://doi.org/10.1016/0888-613X(92)90027-W.

Fagin, R. and J.Y. Halpern. 1991. "A New Approach to Updating Beliefs." In *Uncertainty in Artificial Intelligence*, edited by P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, 347–374. New York: Elsevier.

Fink, P., E. Endres, and M. Schmoll. 2019. *impimp: Imprecise Imputation for Statistical Matching*. https://CRAN.R-project.org/package=impimp. (accessed July 2019).

Joenssen, D.W.H. 2015. *Hot-Deck-Verfahren zur Imputation fehlender Daten – Auswirkungen des Donor-Limits [Hot-Deck Procedures for the Imputation of Missing Data: Effects of the Donor Limit, translation by the authors]*. Ph. D. thesis, Technische Universität Ilmenau. Available at: https://www.db-thueringen.de/receive/dbt_mods_00026076. (accessed July 2019).

Kim, J.K. and W. Fuller. 2004. "Fractional Hot Deck Imputation." *Biometrika* 91: 559–578. Doi: https://doi.org/10.1093/biomet/91.3.559.

Lin, J. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37: 145–151. Doi: https://doi.org/10.1109/18.61115.

Little, R.J.A. and D.B. Rubin. 2002. *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley. Doi: https://doi.org/10.1002/9781119013563.

Manski, C.F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, C.F. 2007. *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

Miranda, E., I. Couso, and P. Gil. 2010. "Approximations of Upper and Lower Probabilities By Measurable Selections." *Information Sciences* 180: 1407–1417. Doi: https://doi.org/10.1016/j.ins.2009.12.005.

Nguyen, H.T. 1978. "On Random Sets and Belief Functions." *Journal of Mathematical Analysis and Applications* 65: 531–542. Doi: https://doi.org/10.1016/0022-247X(78)90161-0.

Nguyen, H.T. 2006. *An Introduction to Random Sets*. Boca Raton: Chapman & Hall/CRC.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/. (accessed July 2019).

Ramoni, M. and P. Sebastiani. 2001. "Robust Learning with Missing Data." *Machine Learning* 45: 147–170. Doi: https://doi.Org/10.1023/A:1010968702992.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer.

Serafino, P. and R. Tonkin. 2017. "Statistical Matching of European Union Statistics on Income and Living Conditions (EU-SILC) and the Household Budget Survey." In *Eurostat: Statistical Working Papers*. Luxembourg: Publications Office of the European Union. Doi: https://doi.org/10.2785/933460.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Vantaggi, B. 2008. "Statistical Matching of Multiple Sources: A Look Through Coherence." *International Journal of Approximate Reasoning* 49: 701–711. Doi: https://doi.org/10.1016/j.ijar.2008.07.005.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Yang, S. and J.K. Kim. 2016. "Fractional Imputation in Survey Sampling: A Comparative Review." *Statistical Science* 31: 415–432. Doi: https://doi.org/10.1214/16-STS569.

Received May 2018
Revised December 2018
Accepted April 2019