LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Eva Endres, Paul Fink, Thomas Augustin

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

Technical Report Number 214, 2018
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

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

Eva Endres* Paul Fink† Thomas Augustin‡
Department of Statistics, LMU Munich

1st March 2018

Abstract

We develop the first statistical matching micro approach reflecting the natural uncertainty arising during the integration of categorical data. A complete synthetic file is obtained by imprecise imputation, replacing missing entries by *sets* of suitable values. We discuss three imprecise imputation strategies and raise ideas on potential refinements by logical constraints or likelihood-based arguments. Additionally, we show how imprecise imputation can be embedded into the theory of finite random sets, providing tight lower and upper bounds for parameters. Our simulation results corroborate that their narrowness is practically relevant and that they almost always cover the true parameters.

Keywords: statistical matching; data integration; imprecise imputation; micro approach; finite random sets; (partial) identification; hot deck imputation;

1 Introduction

Nowadays, a large amount of data is accessible, provided by researchers, companies, or governments. Thus, instead of collecting new data to answer research questions, it is a more convenient alternative to use already available data sources. However, often there is no single data source which includes all information of interest. Statistical matching furnishes a method with which researchers can integrate data collected in different surveys.

Assume that we are interested in three blocks of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, while there are two data files A and B available. Data file A contains n_{A} observations of $(\mathbf{X}, \mathbf{Y})$, and data file B contains n_{B} observations of $(\mathbf{X}, \mathbf{Z})$. 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 not jointly observed variables, is twofold (e.g. D'Orazio et al., 2006b, p.2):

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

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

*eva.endres@stat.uni-muenchen.de
†paul.fink@stat.uni-muenchen.de
‡augustin@stat.uni-muenchen.de

As the schematic representation in Figure 1 suggests, statistical matching can be interpreted as a missing data problem. The observations of the *specific variables* $\mathbf{Y}$ and $\mathbf{Z}$ are missing in a special block-wise pattern in $\mathsf{A} \cup\!\!\!\cup \mathsf{B}$, which denotes the union of the two available data files.[1] This absence of joint information on all variables results in a severe identification problem: the parameters which concern the relationship between $\mathbf{Y}$ and $\mathbf{Z}$ are not directly estimable from $\mathsf{A} \cup\!\!\!\cup \mathsf{B}$ (in the sense of providing a single-valued estimate, see below.)

For instance, D'Orazio et al. (2006b) show different ways to remedy this problem. On basis of their underlying concepts, they can be allocated into three basic groups:
Approaches which

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

(ii) require auxiliary information in terms of a third file or other external information about inestimable parameters,

(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 $\mathbf{Y}$ and $\mathbf{Z}$.[2]

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, p.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. Moreover, auxiliary information, which is the basis of the second group of approaches, is often not available for a certain statistical matching task. Hence, the application of statistical matching taking the underlying uncertainty credibly into account is the means of choice.

In the context of statistical matching, typically the term *uncertainty* refers exclusively to the previously mentioned identification problem.[3] It points to the fact that even if we have complete information on the marginal distributions of $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Z})$, the joint distribution of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ cannot uniquely be determined (e.g. D'Orazio et al., 2006a). Thus, lower and upper bounds on the parameters are the best which can be obtained without relying on strong untestable assumptions or external information. The elaboration of the concept of uncertainty and how to measure it formed the central focus of the papers by Conti et al. (2012) and Conti et al. (2017). Much of the current literature on uncertainty regarding the statistical matching task pays attention to the continuous case, especially on 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), or Di Zio and Vantaggi (2017) deal with statistical matching of categorical data considering different circumstances.

[1] The missingness is induced by the given attribution to a certain data file, and the missing data mechanism in the framework of statistical matching can convincingly be assumed to be missing completely at random (e.g. D'Orazio et al., 2006b, p.6).

[2] These estimates can be interpreted as set-valued point estimates, not to be confused with confidence regions.

[3] Also in this paper, the component of general uncertainty which regards to the sampling process is not addressed.

As emphasized by Conti et al. (2012, p.70), the "third group of techniques" reflecting the natural uncertainty of statistical matching, does usually not "directly aim at reconstructing a complete data set". In the present paper, we introduce imprecise (single) imputation as the first micro approach for categorical data which 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[4] will allow us to derive lower and upper bounds for the parameters of interest.

The paper is structured as follows. Section 2 recalls the background of our work by giving a brief overview on the basic setting of statistical matching, its interpretation as a missing data problem, and hot deck imputation in this framework. 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 set. Section 5 sketches some aspects of refining imprecise imputation in the presence of contextual information or by likelihood-based arguments. After the simulation study in Section 6, we conclude with a summary and outlook in Section 7.

2 Statistical matching

2.1 The basic setting and its missing data interpretation

Throughout the paper, let us assume that we have two data files A and B, indexed by $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$, respectively[5], with n_A and n_B disjoint observation units. Furthermore, let $\mathbf{X} = (X_1, \ldots, X_p)$ be the vector of *common variables*, and $\mathbf{Y} = (Y_1, \ldots, Y_q)$ and $\mathbf{Z} = (Z_1, \ldots, Z_r)$ be the vectors of *specific variables*. Denote the domains of the potential values of X_ℓ, $\ell = 1, \ldots, p$, by $\mathcal{X}_\ell$, 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 the schematic representation in Figure 1, data file A contains exclusively information on $(\mathbf{X}, \mathbf{Y})$, while data file B comprises information on $(\mathbf{X}, \mathbf{Z})$ only. Consequently, there is no observation that contains simultaneous information on $\mathbf{Y}$ and $\mathbf{Z}$. 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 (cf. Figure 1) with $n := n_\mathsf{A} + n_\mathsf{B}$ observations and 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 $P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})$, where $(\mathbf{x}, \mathbf{y}, \mathbf{z}) := (x_1, \ldots, x_p, y_1, \ldots, y_q, z_1, \ldots, z_r)$ depicts the realisations of the variables. 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 $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$.

As previously mentioned, statistical matching may be regarded as missing data problem. Hence, a natural strategy to solve the statistical matching task is imputation, i.e. the substitution of the missing entries with suitable/similar real or artificial values to derive a complete (but partially synthetic) data set. To prepare our method, we focus in the

[4] See, for instance, Nguyen (2006) or Couso et al. (2014).

[5] Without loss of generality, we assume for convenience that the index sets are disjoint, e.g. $\mathcal{I}_\mathsf{A} = \{1, \ldots, n_\mathsf{A}\}$ and $\mathcal{I}_\mathsf{B} = \{n_\mathsf{A} + 1, \ldots, n_\mathsf{A} + n_\mathsf{B}\}$.

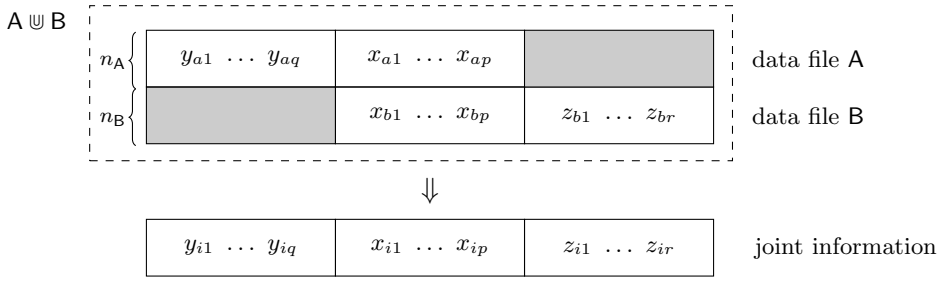

Figure 1: Schematic representation of the statistical matching problem (cf. D'Orazio et al., 2006b, p.5 (modified)).

following section 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.[6] By applying hot deck imputation, we ensure that only so-called *live* values, i.e. actually observed and no artificial values are substituted, and 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 non-parametric (e.g. Andridge and Little, 2010).

2.2 Hot deck imputation for statistical matching

In the context of statistical matching, hot deck imputation belongs to the group of non-parametric micro approaches. D'Orazio et al. (2006b, Chapter 2.4) describe it as follows for four variables (X_1, X_2, Y_1, Z_1). The data samples A and B are assigned to the roles of *recipient file* and *donor file*[7]. Since it is a symmetric problem, they only describe the case where A is the recipient file and B the donor file. The reverse case works analogously.

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. This means that every missing entry of the specific variable Z_1 in the recipient file A, i.e. z_{1a}, $a \in \mathcal{I}_\mathsf{A}$, is replaced by the synthetic value $\tilde{z}_{1a} := z_{1b}$, $b \in \mathcal{I}_\mathsf{B}$, where b is the randomly chosen observation unit from the index set $\mathcal{I}_\mathsf{B}$ of data file B, and hence $\tilde{z}_a \in \{z_1, \dots, z_{n_\mathsf{B}}\}$. Hence, the a-th observation of data file A is composed of $(x_{a1}, x_{a2}, y_{1a}, \tilde{z}_{1a})$, 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 to be selected. Thus, this procedure implicitly induces the independence of the common variables and the specific variables.

A more promising procedure is the assignment of donor and recipient records within groups of similar (homogeneous) records which are developed by exploiting the information of the common variables[8]. The instantiations of selected categorical common variables[9] are used to generate groups of similar records in both the recipient and the donor file.

[6] For the general missing data case, see e.g. Little and Rubin (2002, p.66).

[7] The choice of whether only A, only B, or $\mathsf{A} \cup \mathsf{B}$ should be imputed depends on many factors. In this paper, we impute $\mathsf{A} \cup \mathsf{B}$ without loss of generality. See, for instance, D'Orazio et al. (2006b, pp.35–36) for a discussion on this issue.

[8] The choice of the common variables which are actually used to perform statistical matching (the so-called *matching variables*) is of high impact on the resulting matching quality. It is desirable that the common are highly correlated with, or good predictors for the specific variables (Rässler, 2002, p.10).

[9] See, for instance, D'Orazio et al. (2017) on how to choose the *matching variables*.

Following D'Orazio et al. (2006b), we call these groups *donation classes*[10].

Consider again data file A as the recipient file. The first step is the assignment of all observations in $\mathsf{A} \Cup \mathsf{B}$ to donation classes. For this purpose, 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 $\mathbf{x}$. 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 of the specific variable Z_1 of a observation unit from A in the d-th donation class, i.e. z_{1a}, $a \in \mathcal{I}^d_{\mathsf{A}}$, is replaced by $\tilde{z}_{1a} := z_{1b}$, $b \in \mathcal{I}^d_{\mathsf{B}}$, which is the corresponding value of a randomly chosen observation from the donation class $\mathcal{I}^d_{\mathsf{B}}$, and hence $\tilde{z}_a \in \{z_b : b \in \mathcal{I}^d_{\mathsf{B}}\}$ for all $a \in \mathcal{I}^d_{\mathsf{A}}$.

Using donation classes, the imputation of $\mathbf{Z}$ is conditional on $\mathbf{X}$, and thus basically reproducing the empirical conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ in A. Since no common observations of all variables are available, an additional conditioning on $\mathbf{Y}$ is not possible, which means that, in principle, empirical conditional independence given $\mathbf{X}$ of the imputed values of $\mathbf{Y}$ and the values of $\mathbf{Z}$ is implicitly established (cf. Rässler, 2002, pp.200 – 204).[11]

Any synthetic data set with observations $(\mathbf{x}_a, \mathbf{y}_a, \tilde{\mathbf{z}}_a)_{a \in \mathcal{I}_{\mathsf{A}}}$ and $(\mathbf{x}_b, \tilde{\mathbf{y}}_b, \mathbf{z}_b)_{b \in \mathcal{I}_{\mathsf{B}}}$ naturally delivers estimates of the underlying joint distribution by evaluating the observed relative frequencies. For any event $E = E_{\mathcal{X}} \times E_{\mathcal{Y}} \times E_{\mathcal{Z}}$ with $E_{\mathcal{X}} \subseteq \mathcal{X}$, $E_{\mathcal{Y}} \subseteq \mathcal{Y}$ and $E_{\mathcal{Z}} \subseteq \mathcal{Z}$, one obtains

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

In the context of missing data, it is a well-known problem that single imputations are not able to reflect the uncertainty which arises from the missingness. Therefore, it is commonly recommended to apply *multiple imputation* techniques (e.g. Little and Rubin, 2002, Chapter 5.4), where the replacement of the 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 valid point estimates. Such multiple imputation techniques have been further developed by (Rässler, 2002, Chapter 4) for the application in the context of 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, p.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 now develop the concept of imprecise imputation where we suggest to impute a *set* of plausible values for a missing entry. This leads to precise

[10] Little and Rubin (2002) call these groups *adjustment cells*.

[11] Another consequence of an imputation process based on donation classes is that the observations are slightly dependent of each other. In accordance with most literature on statistical matching, this aspect is also not problematised here and in the following sections.

observations ($\mathbf{x}$, $\mathbf{y}$) in A and ($\mathbf{x}$, $\mathbf{z}$) in B, and to *imprecise*, i.e. set-valued, observations $\tilde{\mathfrak{z}}$ in A, and $\tilde{\mathfrak{y}}$ in B.

The following subsections detail and illustrate imprecise imputation. Depending on how strong and trustworthy the underlying relationship between the common and specific variables is, three different ways of determining the sets of plausible values to be imputed are introduced. Without loss of generality, let again A be the recipient and B the donor file, and let the donor classes be defined in the same way as in Section 2.2.

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

$$\tilde{\mathfrak{z}}_{a\ell} := \mathcal{Z}_\ell\,, \qquad \forall a \in \mathcal{I}_\mathsf{A}\,,\ \ell = 1, \ldots, r\,. \tag{2}$$

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

$$\tilde{\mathfrak{z}}_{a\ell} := \left\{z_{b\ell} : b \in \mathcal{I}^d_\mathsf{B}\right\}, \qquad \forall a \in \mathcal{I}^d_\mathsf{A}\,,\ d = 1, \ldots, D,\ \ell = 1, \ldots, r\,. \tag{3}$$

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

$$\widetilde{\mathfrak{z}}_a := \left\{(z_{b1}, \ldots, z_{br}) : b \in \mathcal{I}^d_\mathsf{B}\right\}, \qquad \forall a \in \mathcal{I}^d_\mathsf{A}\,,\ d = 1, \ldots, D\,. \tag{4}$$

3.2 Illustration and discussion of the different types of imprecise imputation

3.2.1 Domain imputation

The most cautious way to determine the set of plausible values which 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 $z_{a\ell}$, $a \in \mathcal{I}_\mathsf{A}$, $\ell = 1, \ldots, r$, is substituted by the set of all possible realisations of Z_ℓ, i.e. its domain $\mathcal{Z}_\ell$. Hence, $\tilde{\mathfrak{z}}_{a\ell} := \mathcal{Z}_\ell$, $\forall a \in \mathcal{I}_\mathsf{A}$, becomes a set-valued entry in data file A, where all elements of the set are treated as equally plausible. the imputed sets for one variable are equal for all observations. This procedure is briefly illustrated in the following 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 set.*

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 line. The numbers in curly brackets depict the sets of possible realisations of the corresponding variables, i.e. 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.

As previously mentioned, domain imputation is very cautious, and it thus can also be applied if the common variables are not good predictors for the specific variables. Applying this imputation approach, it is guaranteed that the true (but missing) value is always element of the imputed set. As it neglects any available dependence structure between the common and specific variables, we will introduce two other methods to determine the set of values for imputation.

3.2.2 Variable-wise imputation

If the common variables are good predictors for the specific variables, domain imputation nevertheless ignores these relationships and alleviates existing dependencies. The imputation of only live values within donation classes ensures that the 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 interests become more narrow. If $q \geq 2$ or $r \geq 2$, with due regard of 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 the two blocks of specific variables simultaneously (cf. e.g. Joenssen, 2014, Chap. 3).

Without loss of generality, let again 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 $\mathbf{X}$, following the notation as introduced in Section 2.2. For every observation $a \in \mathcal{I}_\mathsf{A}^d$, the missing entry $z_{a\ell}$ of the variable Z_ℓ, $\ell = 1, \ldots, r$, is substituted by the set of all live values of this variable from the same donation class in the donor file B, resulting in (3).

Minimal Example 2. *Consider the same data situation as in Example 1. Now we illustrate the application of the just described variable-wise imputation. The grey background displays the different donation classes based on the combinations of the realisations of X_1 and X_2, both of which are used as matching variables.*

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 variables and the specific variables, however, the successive imputation of single variables breaks the dependence structure among the specific variables. Little and Rubin (cf. 2002, p.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.

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 $\widetilde{\mathfrak{z}}_a$, which have been observed in the donor file B and the same donation class d , as in (4) . This strategy ensures that the dependencies among the specific variables $\mathbf{Z}$ remain unchanged. The following example illustrates the simultaneous imputation procedure.

Minimal Example 3. *Consider again the starting point as in Example 1. 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}), $b = 3, 4, 5$.*

(y_1, y_2)	x_1	x_2	(z_1, z_2)
$(\mathbf{1}, \mathbf{2})$	**1**	**0**	$\{(0,0);(1,1)\}$
$(\mathbf{0}, \mathbf{2})$	**0**	**0**	$\{(1,2)\}$
$\{(1,2)\}$	**1**	**0**	$(\mathbf{0}, \mathbf{0})$
$\{(1,2)\}$	**1**	**0**	$(\mathbf{1}, \mathbf{1})$
$\{(0,2)\}$	**0**	**0**	$(\mathbf{1}, \mathbf{2})$

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 recommend to impute the domains $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$ or the Cartesian product of the domains $\mathcal{Z}$, respectively.

Variable-wise and case-wise imputation are a set-valued generalisation of hot deck imputation based on homogeneous donation classes. They transfer, to a different extent, the association structure between the common and the observed specific variables into the synthetic file.

So far, we have produced synthetic data files with the aid of imprecise imputation. In contrast to widely-adopted imputation procedures yielding singe-valued data, we are here in the situation of statistical analysis of partially set-valued data. To frame it formally, imprecise imputation will be embedded into the concept of finite disjunctive random sets, which will allow 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 one taken by the case-wise imputation. The imputation results of the other procedures can by transferred by taking the Cartesian product:

$$\widetilde{\mathfrak{z}}_a = \tilde{\mathfrak{z}}_{a1} \times \ldots \times \tilde{\mathfrak{z}}_{ar} \ .$$

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 the situation probabilistically. For this purpose,

the direct formalisation in terms of the vectors $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ of random variables[12] and corresponding realisations is no longer sufficient. Two types of generalisations, which indeed will 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.

In this section it will be shown how set-valued observations, and thus in particular the resulting data files of the three imprecise imputation procedures, are covered by the concept of *disjunctive random sets*, also known as *ill-perceived random variables* (Couso et al., 2014)[13]. 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.[14]

4.1 Random set formulation of imprecise imputation

The true random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ map from the underlying population space, denoted by Ω in the sequel, into the domains $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, yielding realisations $\mathbf{x}, \mathbf{y}, \mathbf{z}$, respectively. Now, either $\mathbf{y}$ or $\mathbf{z}$ are not available, but are replaced by $\tilde{\mathfrak{y}}$ and $\tilde{\mathfrak{z}}$, respectively, according to Equations (2), (3) and (4), depending on the chosen imprecise imputation procedure. To formalise this situation, we follow the common practice in statistical matching to take a conditional perspective on the sampling process, treating $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$ as fixed.

This allows us to globally replace $\mathbf{Y}$ and $\mathbf{Z}$ by the set-valued variables $\mathfrak{Y}$ and $\mathfrak{Z}$ (with realisations $\mathfrak{y}$ and $\mathfrak{z}$). The imputed values are already sets, so they fit in nicely, but in order to deal with the already observed realisation, we regard them as singletons containing just the observed value, e.g. $\mathfrak{z}_{b\ell} = \{z_{b\ell}\},\ \forall b \in \mathcal{I}_\mathsf{B}, \ell = 1, \ldots, r$. 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 Γ and define $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then, with $\mathcal{X} \times 2^\mathcal{Y} \times 2^\mathcal{Z} \subset 2^\mathcal{W}$,

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

is a finite non-empty random set (cf. Definition 3.1 in Nguyen, 2006, p.35), satisfying the required measurability condition by equipping $2^\mathcal{W}$ 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 so-called disjunctive way (cf., e.g. Couso et al., 2014).[15]

In general, any disjunctive random set Γ induces an upper inverse Γ^* and a lower inverse Γ_*. When considering an event of interest $E \subseteq \mathcal{W}$, 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 E, while the lower inverse contains only those elements of the

[12] Throughout this paper we use the term *random variable* to refer to a mapping to the real numbers as well as to some non-numerical finite space. In context of the latter the term *random element* is sometimes used for the sake of distinction.

[13] See also in particular Nguyen (2006)

[14] The interpretation of the set-valued quantities as disjunctive random sets corresponds to the view of Dempster (1967), on which the so-called Dempster-Shafer theory of belief functions (Shafer, 1976) is built, which has become very popular in artificial intelligence (See, e.g. Denœux, 2016). Yet it comes with different interpretations of derived concepts, especially when considering conditioning. Nevertheless, many technical results can be used.

[15] See also the discussion in Couso and Dubois (2014).

population whose (non-empty) image is entirely contained within E:[16]

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

and

$$\Gamma_*(E) := \big\{\omega \in \Omega : \Gamma(\omega) \subseteq E\big\} \,. \tag{7}$$

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:[17]

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

If we turn to the view of an underlying, ill-perceived random variable $W_0 : \Omega \longrightarrow \mathcal{W}$, only knowing that the unobserved true value $W_0(\omega)$ lies within the observed set $\Gamma(\omega)$ (with probability one) , it can be shown (cf., e.g. Couso et al., 2014) that for every event E in $\mathcal{W}$ the upper and lower probabilities induced by the random set enclose the probability of W_0:

$$P_*(E) \leq P_{W_0}(E) \leq P^*(E) \quad \forall E \subseteq \mathcal{W} \,.$$

This leads to another way to interpret a random set, namely as producing a family of compatible precise probability measures $\mathcal{P}(\Gamma)$, which is a subset of the set $\mathcal{P}$ of all probability measures on $(2^{\mathcal{W}}, 2^{2^{\mathcal{W}}})$. In the present special case of finite $\mathcal{W}$, the set $\mathcal{P}(\Gamma)$ coincides with the credal set $\mathcal{M}(P^*)$, i.e. those precise probability measures that respect the upper and lower bounds defined by P^* and P_* event-wise (cf. Miranda et al., 2010)[18], 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 just the envelopes of all probability measures P in $\mathcal{M}(P^*)$

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

Indeed, P^*, P_* and $\mathcal{M}(P^*)$ are three mathematically equivalent formulations, which can be transferred into each other. Therefore, from a 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_{W_0}, our embedding into random sets provides us with a set of distributions $\mathcal{M}(P^*)$ induced by P_{W_0} and the concretely chosen imputation procedure such that $\mathcal{M}(P^*)$ contains P_{W_0}. By construction, this is the smallest set that is deducible from the concrete imputation procedure without adding further assumptions/knowledge. Dually, $P^*(E)$ and $P_*(E)$ are the narrowest bounds, deducible on the probabilities of an event E.

[16] In a heuristic formulation the upper inverse looks at all aspects that do not contradict E, while the lower inverse collects all aspects that necessarily imply E.

[17] In order to improve readability we do not mark the image probability measure induced by the random set Γ, i.e. $P_\Gamma = P$, and we proceed analogously with the corresponding set-functions P^* and P_*. If we refer to a different image measure, the according inducing random quantity will be set as subscript to P.

[18] Nguyen (1978) showed that if $\mathcal{W}$ is finite, the probability distribution induced by Γ corresponds to the basic probability assignment in Dempster-Shafer theory and thus makes the belief function mathematically equivalent to P_*, and thus technical results from that area may be used as well.

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 $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ are eventually 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 $P(C) > 0$ for the conditioning event C. One obtains the following closed form results[19] for the upper conditional probability[20]

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

and the lower conditional probability

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

However, literature about imprecise probabilities warns that special care needs to be taken when performing conditioning. A distinction of the nature of conditioning has to be made, resulting in several concepts which in the classical setting of precise probabilities lead to the same numerical results.[21]

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 corresponding to the random set need to be estimated from a finite sample. Consequently, we take our synthetic data set derived from imprecise imputation as consisting of n realisations γ_i, $i \in \mathcal{I}$, of the corresponding generic random set Γ from (5).[22] Referring to (8) with (6) and (7), we obtain, in generalisation of (1), for our generic event $E = E_{\mathcal{X}} \times E_{\mathcal{Y}} \times E_{\mathcal{Z}}$

[19] The second equality in (9) and (10) is a consequence of the so-called two-monotonicity of P_* from (8). The interested reader is referred to, e.g. de Campos et al. (1990), Couso et al. (2014) and Fagin and Halpern (1991), where (9) and (10) are derived.

[20] $\bar{S}$ denotes the complement of S.

[21] The approach just introduced in (9) and (10) can be rigorously justified in Walley's framework of coherent inference (cf. Walley, 1991, Chapter 6). As proven by Jaffray (1992) in the context of capacities, those upper and lower probabilities coincide in the theory of belief functions with plausibility and belief functions, respectively. However, they are to be interpreted in the notion of 'focusing', also known as 'Fagin-Halpern updating'. The second major concept of defining a conditional imprecise probability, the revision of the probability distribution, will lead to lower and upper probabilities, which numerically coincide with belief and plausibility if they are obtained via so-called Dempster's rule of conditioning. (See Dubois and Prade (1992) and Fagin and Halpern (1991) for a comparison of both concepts.)

[22] In the approach within the framework of belief functions, leading to numerically identical estimators, the basic probability assignment is set to the relative frequencies, and then belief and plausibility are derived for all the events of interest.

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

and

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

From that also an estimate of the induced underlying set of probability measures can be derived as

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

For comparing the estimates resulting from the different types of imputation procedures, it is essential to recall that, by construction, the generated set-valued data sets are ordered by set inclusion, with respect to all compatible underlying precise data sets. The set resulting from domain imputation is a (non-strict) 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 Section 3.1, it holds naturally that

$$\widehat{\mathcal{M}}\Big(P^{*^{CW}}\Big) \subseteq \widehat{\mathcal{M}}\Big(P^{*^{VW}}\Big) \subseteq \widehat{\mathcal{M}}\Big(P^{*^{D}}\Big) \qquad (14)$$

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

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

This allows to compare the results obtained by 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 it can be seen from the relations between the different sets of probabilities in Equation (14), the one induced by case-wise imputation can be regarded as containing probability distributions neighbouring the one under conditional independence. The others can be interpreted to deviate even more from conditional independence, with domain imputation as approach demonstrable neglecting any conditional dependence structure in the construction of its bounds. For domain imputation the bounds are maximal (but not vacuous), despite constricting the parameter space.

Minimal Example 4. *For demonstrative purpose 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 Eq.* (9)*. We estimate the upper joint probability with Eq.* (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 Eq.* (12) *by counting how many observations have only 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 Eq.* (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* $[\frac{1}{3}; \frac{1}{2}]$*.*

5 Refining imprecise imputation: logical constraints and likelihood-based arguments

During the imputation process, it is possible that imprecise variable-wise, and in particular, domain imputation creates combinations of variable realisations which are contextually unjustified. Assume, for instance, that Y_1 indicates sex (0: male/ 1: female), and Z_1 is a binary variable describing the pregnancy state (0: not pregnant/ 1: pregnant) in the previous examples. We can appropriately assume that the combination $(y_1, z_1) = (0, 1)$ is impossible considering usual circumstances and that we want to exclude this combination from our synthetic data. Following D'Orazio et al. (e.g. 2006b), we call such rules *logical constraints*[23].

D'Orazio et al. (e.g. 2006b, p.126) distinguish between two main cases of logical constraints when dealing with categorical data:

(i) *'existence of some quantities'* (which corresponds to the previous example, where pregnant men have to be excluded from the synthetic data), and

(ii) *'inequality constraints'* (e.g. people who eat healthy have a higher probability to have a better body mass index than people who eat fast food regularly).

The first case can easily be incorporated in the imputation step. For that purpose the data file can be transformed into the tuple notation exemplified in Example 1 by applying the Cartesian product to each observation unit.

Minimal Example 5. *Numbers in bold represent again the information of the original data files* A *and* B*.*

$(y_1, y_2, x_1, x_2, z_1, z_2)$
$\{(\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 0, 0); (\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 0, 1); (\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 0, 2); (\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 1, 0); (\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 1, 1); (\mathbf{1}, \mathbf{2}, \mathbf{1}, \mathbf{0}, 1, 2)\}$
$\{(\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 0, 0); (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 0, 1); (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 0, 2);$ ~~$(\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 1, 0); (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 1, 1); (\mathbf{0}, \mathbf{2}, \mathbf{0}, \mathbf{0}, 1, 2)$~~$\}$
$\{(0, 0, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}); (0, 1, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}); (0, 2, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}); (1, 0, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}); (1, 1, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}); (1, 2, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})\}$
$\{$~~$(0, 0, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}); (0, 1, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}); (0, 2, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1});$~~ $(1, 0, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}); (1, 1, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}); (1, 2, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1})\}$
$\{$~~$(0, 0, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2}); (0, 1, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2}); (0, 2, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2});$~~ $(1, 0, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2}); (1, 1, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2}); (1, 2, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{2})\}$

[23]These kinds of zero frequencies which are caused by impossible combinations are also known under the terms *fixed zeros* (cf., e.g. Fienberg, 2007) or *structural zeros* (cf., e.g. Berger and Zhang, 2005).

Using this tuple notation, it is straightforward to identify and remove tuples with entries that contradict these logical constraints. In the example, tuples with the combination $(y_1, z_1) = (0, 1)$ *are crossed out, and the corresponding set-valued variables in Section 4 have to be adapted.*

The second type of logical constraints can be considered in the estimation step if we interpret them as additional constraints on the estimated set $\widehat{\mathcal{M}}(P^*)$ of probabilities derived from our imprecise imputation (cf. Equation (13)). Since, by construction, $\widehat{\mathcal{M}}(P^*)$ can be represented by a convex polyhedron in $\mathbb{R}^{|\mathcal{W}|-1}$, in particular linear constraints can be incorporated very conveniently.

In further extension of the first type of constraints, one could also argue that, in particular in very large data sets, not only contextually impossible but also combinations of values that showed to be very rare should be excluded from the set to be imputed variable- or case-wise. In principle, 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.[24] Going further, confining imprecise imputation to sets $\mathcal{S}_\delta$ of values with relative likelihood exceeding δ would gradually push with increasing δ the cautious perspective taken here into the background. Technically, such approaches can still be handled within the framework developed in Section 4, after appropriately adjusting the multi-valued mapping Γ. With increasing δ the induced set of imputed values would become naturally smaller, until each missing observation is just replaced by the corresponding most frequent observation (if unique). This would lead to fixed single-valued imputation and estimates degenerating to a single value.

6 Simulation study of imprecise imputation

To investigate the quality of imprecise imputation, we have performed a simulation study. For this purpose, we simulated a complete categorical data file $\mathsf{A} \cup \mathsf{B}$ 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 $\mathbf{Z}$ and $\mathbf{Y}$ 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 the lower bound. 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 which covers the component of the true parameter, the better the procedure performs. In the following, we will detail the simulation design, parameters, and results.

6.1 Simulation design

The starting point of our simulation analysis are two categorical data files A and B. Both of them contain information on four common variables $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, and four specific variables $\mathbf{Y} = \{Y_1, Y_2, Y_3, Y_4\}$ or $\mathbf{Z} = \{Z_1, Z_2, Z_3, Z_4\}$, 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:

[24] This idea is motivated by the approach of Cattaneo (2013), who developed a likelihood-based decision theory based on sets of parameter values that are not too implausible given the observed data.

Scenario	C	Jensen-Shannon Divergence
1	$[0; 0.2)$	far away
2	$[0.2; 0.6)$	far away
3	$[0.6; 1)$	far away
4	$[0; 0.2)$	close
5	$[0.2; 0.6)$	close
6	$[0.6; 1)$	close

Table 1: Overview on the simulation scenarios.

1. The strength of the bivariate associations in terms of the corrected contingency coefficient[25] C (low: $C \in [0, 0.2)$, medium: $C \in [0.2, 0.6)$, high: $C \in [0.6, 1)$).

2. The Jensen-Shannon divergence (e.g. Lin, 1991) from the marginal distribution of the common variables to the discrete uniform distribution.

This leads to the six simulation scenarios depicted in Table 1. Additionally, we vary

3. the numbers of observations $n_{\mathsf{A}} = n_{\mathsf{B}} \in \{50, 100, 250\}$, and

4. the dependence structure among the variables (cf. Figure 2).

Altogether, we achieve 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.

6.2 Simulation parameters

As already stated by Rässler (2002, p.10), the common variables should be good predictors for the specific variables. This ensures that the donation classes are suitable to generate homogeneous groups of observations which 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 which are covered by our simulation design. The upper six variables of each design represent the binary variables, 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 first design in the upper left block, 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 χ^2-coefficient for contingency tables but it is corrected for the number of observations as well as the number of categories.

At the first sight, the number of observations plays a counter-intuitive role in this simulation study. We expect that the distances between the lower and upper bounds for the parameters of interest increase in situations with a higher number of observation. This is due to the fact that a growth of the number of observations also causes an increase of 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

[25] Also known as Sakoda's adjusted Pearson's C.

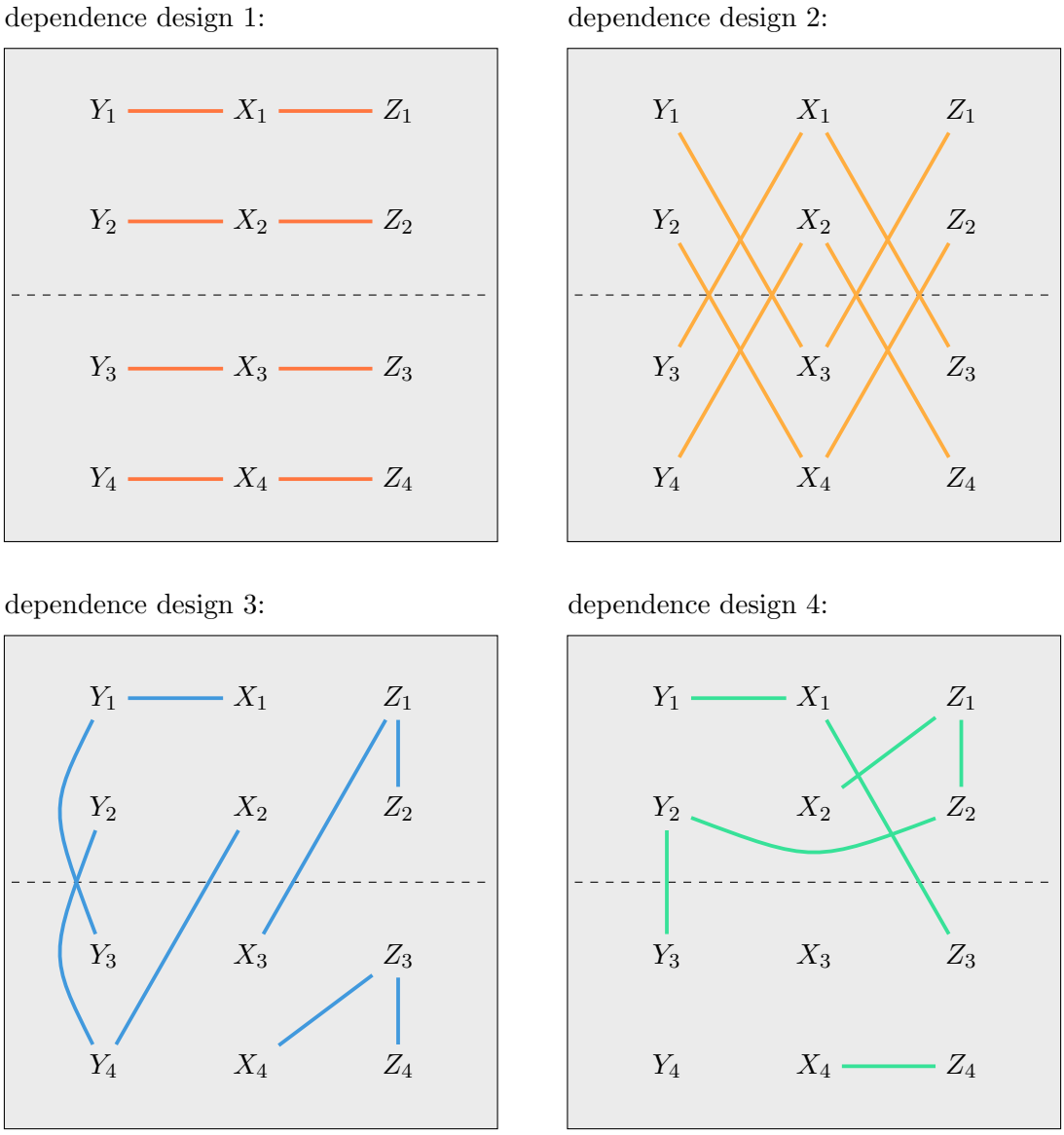

Figure 2: Four different dependence structures among the variables in the simulation study. An edge between two variables displays dependence between them.

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	0.8500	0.9446	0.9725	0.9921	1.0000	0.9649

imputation procedure	min.	1st quartile	median	3rd quartile	max.	mean
domain	1	1	1	1	1	1
variable-wise	0.9994	0.9998	0.9999	0.9999	1.0000	0.9998
case-wise	0.9994	0.9998	0.9999	0.9999	1.0000	0.9998

Table 2: 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.

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 Section 3.2.4, the domain for the missing entries which corresponds to a minimum of information which in turn leads to bounds which are further apart.

6.3 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 which equal the mean distances between the upper and lower bound. An interval width of 0 corresponds to a precise estimation.

Table 2 shows that the true values of the marginal and the joint distributions almost always lie inside the estimated bounds. For domain imputation, the true value was indeed always element of the set of estimated parameters.

The interval width was separately analysed for the components of the marginal distributions and the 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 which 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 less 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 which are close to the uniform distribution and few direct connections between the specific variables and the common variables. For the simulation designs with a higher divergence to the uniform distribution, the variation of the interval widths are 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. This result is somewhat counter-intuitive, 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 less 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.

The results furthermore show that the interval widths of the marginal distributions slightly increase with a growing number of observations. 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 effects 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 the dependence structure among the variables in a data file has, as expected, 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 less connections. The last dependence structure where there are much less 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 which cover the components of the true parameter value almost always. (The number of cases where 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.

7 Concluding remarks

We 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, i.e. 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 in the original files A and B. They also differ in the amount of data constellations produced beyond the ones obtained by single or multiple imputation under the conditional independence assumption.

The most cautious approach, domain imputation, breaks the dependence structure available in the original data during the creation of the synthetic part of the resulting complete file. It is the most general approach in the sense that it does not rely on typical assumptions usually made in the framework of statistical matching. Against this background, domain imputation is even a suitable statistical matching approach when the common variables are no good predictors for the specific variables since it encompasses the results for all possible realities which are compatible with the available data. On the other hand, imprecise imputation based on donation classes is able to utilise even the smallest observed dependencies between the common and the specific variables.

Embedding imprecise imputation into the framework of finite random sets allows to derive set-valued estimates of the underlying true parameters. These estimates – possibly after their refinement by external information, see e.g. Section 5 – reflect the uncertainty, inherent to the identification problem of statistical matching. The estimation procedure utilises to full extent the set-valued information without artificially reducing the complexity of the imputed sets. Simulation results, based on a new simulation technique for dependent categorical data, corroborated 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 an strongly unbalanced statistical matching situation where, e.g. $n_{\mathsf{A}} \ll n_{\mathsf{B}}$, imprecise imputation can be applied straightforward 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 $\mathbf{Y}$ are precise.

Moreover, the imprecise imputed data set with synthetic set-valued observations can be used as a starting point to derive one or multiple data sets 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 on those data sets by common multiple imputation techniques. However, then one would loose to a considerable extent sight of the conviction of this work 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 and on the other hand, the performance of imprecise imputation applied to real data should be assessed in detail. A further natural progression of this work is the comparison to existing statistical matching macro approaches which 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, and 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, worthwhile to be elaborated and evaluated in detail.

Acknowledgements

The authors are grateful to Christoph Jansen and Georg Schollmeyer for very stimulating discussions. The first author also thanks the LMUMentoring programme for versatile support.

References

Ahfock, D., Pyne, S., Lee, S. X. and McLachlan, G. J. (2016). Partial identification in the statistical matching problem, *Computational Statistics & Data Analysis* **104**: 79–90.

Andridge, R. and Little, R. (2010). A review of hot deck imputation for survey non-response, *International Statistical Review* **78**(1): 40–64.

Augustin, T., Coolen, F. P. A., de Cooman, G. and Troffaes, M. C. M. (eds) (2014). *Introduction to Imprecise Probabilities*, Wiley, Chichester.

Barbiero, A. and Ferrari, P. A. (2017). An R package for the simulation of correlated discrete variables, *Communications in Statistics – Simulation and Computation* **46**(7): 5123–5140.

Berger, V. and Zhang, J. (2005). Structural zeros, *in* B. Everitt and D. Howell (eds), *Encyclopedia of Statistics in Behavioral Science*, Wiley, Chichester, pp. 1958–1959.

Cattaneo, M. (2013). Likelihood decision functions, *Electronic Journal of Statistics* **7**: 2924–2946.

Conti, P. L., Marella, D. and Scanu, M. (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators, *Computational Statistics & Data Analysis* **53**(2): 354–365.

Conti, P. L., Marella, D. and Scanu, M. (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**(2): 967–94.

Conti, P., Marella, D. and Scanu, M. (2012). Uncertainty analysis in statistical matching, *Journal of Official Statistics* **28**(1): 69–88.

Couso, I. and Dubois, D. (2014). Statistical reasoning with set-valued information: Ontic vs. epistemic views, *International Journal of Approximate Reasoning* **55**(7): 1502–1518.

Couso, I., Dubois, D. and Sánchez, L. (2014). *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*, Springer, Cham.

de Campos, L. M., Lamata, M. T. and Moral, S. (1990). The concept of conditional fuzzy measure, *International Journal of Intelligent Systems* **5**(3): 237–246.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping, *The Annals of Mathematical Statistics* **38**(2): 325–339.

Denœux, T. (2016). 40 years of Dempster-Shafer theory, *International Journal of Approximate Reasoning* **79**: 1–6.

Di Zio, M. and Vantaggi, B. (2017). Partial identification in statistical matching with misclassification, *International Journal of Approximate Reasoning* **82**: 227–241.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006a). Statistical matching for categorical data: Displaying uncertainty and using logical constraints, *Journal of Official Statistics* **22**(1): 137–157.

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

D'Orazio, M., Di Zio, M. and Scanu, M. (2017). The use of uncertainty to choose matching variables in statistical matching, *International Journal of Approximate Reasoning* **90**: 433–440.

Dubois, D. and Prade, H. (1992). Evidence, knowledge, and belief functions, *International Journal of Approximate Reasoning* **6**(3): 295–319.

Fagin, R. and Halpern, J. Y. (1991). A new approach to updating beliefs, *in* P. Bonissone, M. Henrion, L. Kanal and J. Lemmer (eds), *Uncertainty in Artificial Intelligence, UAI'91*, Elsevier Science, New York, pp. 347–374.

Fienberg, S. (2007). *The Analysis of Cross-Classified Categorical Data*, 2nd edn, Springer, New York.

Jaffray, J. Y. (1992). Bayesian updating and belief functions, *IEEE Transactions on Systems, Man, and Cybernetics* **22**(5): 1144–1152.

Joenssen, D. W. H. (2014). *Hot-Deck-Verfahren zur Imputation fehlender Daten – Auswirkungen des Donor-Limits*, PhD thesis, Technische Universität Ilmenau.

Lin, J. (1991). Divergence measures based on the Shannon entropy, *IEEE Transactions on Information Theory* **37**(1): 145–151.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd edn, Wiley, Hoboken, NJ.

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

Manski, C. (2007). *Identification for Prediction and Decision*, Harvard University Press, Cambridge.

Miranda, E., Couso, I. and Gil, P. (2010). Approximations of upper and lower probabilities by measurable selections, *Information Sciences* **180**(8): 1407–1417.

Nguyen, H. T. (1978). On random sets and belief functions, *Journal of Mathematical Analysis and Applications* **65**(3): 531–542.

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

Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data, *Machine Learning* **45**(2): 147–170.

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

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

Vantaggi, B. (2008). Statistical matching of multiple sources: A look through coherence, *International Journal of Approximate Reasoning* **49**(3): 701–711.

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

Appendices

A Why we need a new simulation procedure

To generate simulated categorical data meeting all the desired properties, we propose a new procedure which we detail in the following. But, as a start, we want to elucidate why conventional simulation approaches are not suitable for our requirements. The key aspects can be 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 and which 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 which fulfils all the desired properties. Furthermore, we would argue that it is necessary to consider several joint distributions 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.

(iii) The simulation of categorical data which imply a certain dependence structure can also be realised using a probabilistic graphical model like a Bayesian network. The major problem with this way to proceed is the resulting conditional independence among parts of our variables. If the – in real-world applications usually 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 which are based on a Gaussian copulas like the one suggested by Barbiero and Ferrari (2017).

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

B Simulation procedure

For this purpose, we created a new simulation procedure which is directly based on tables of relative frequencies and a suitable association measure.[26] The bivariate associations within the simulated data can be expressed by this association measure on bivariate frequency tables of sizes 2×2, 2×3, and 3×3 (c.f. the domains listed in Section 6).

[26] As mentioned above, we use the corrected contingency coefficient to express the strength of associations. Since 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 which represents the set of all possible frequency tables of above mentioned sizes. S is created by taking all combinations of two discrete probability (marginal) distributions, whose event probabilities are strictly positive[27] and on a one percent grid. Thus, S covers a large variety of marginal distributions and association measures ($|S| = 48\,044\,502$).

In a second step, we randomly draw one frequency table from $S^\star$ for each bivariate association depicted in Figure 2, where $S^\star \subseteq S$ denotes the set of probability tables which 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 $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. To meet the challenges of a statistical matching framework, we split this data file into two files A and B, with $n_\mathsf{A} = n_\mathsf{B}$, and remove the observations $\mathbf{z}$ from A and $\mathbf{y}$ from B, respectively.

C Simulation results

Figures 3 – 8 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 (cf. 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 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 better visible, the values itself are not transformed.

[27] Zero-entries in the marginal distributions lead to zero-entries in the table under independence. This entails that the χ^2-coefficient and all association measures based on it are not defined.

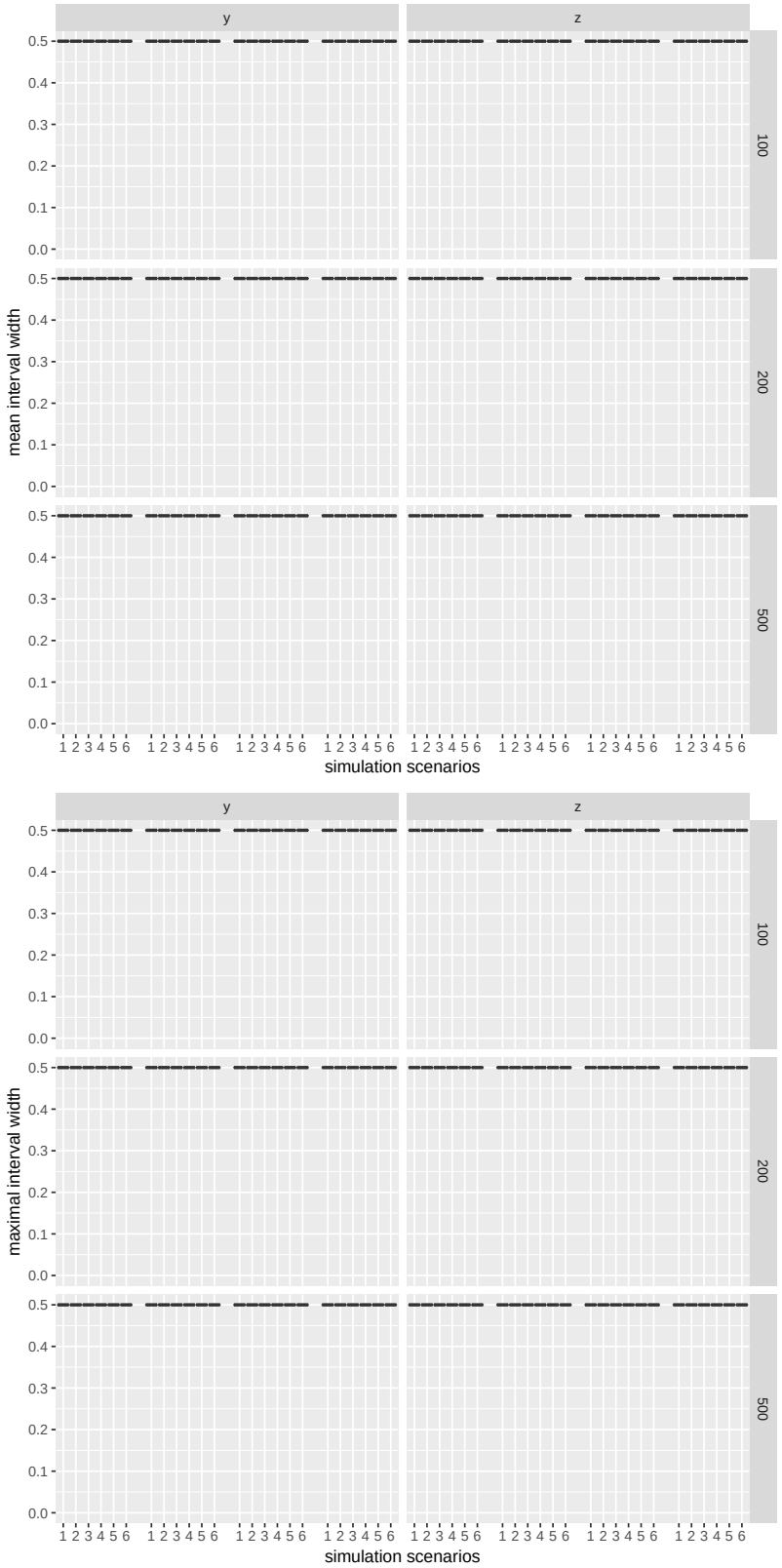

Figure 3: Mean and maximal interval widths of the components of the marginal distributions of the specific variables for domain imputation. The two columns display the pooled results for the marginals of the specific variables $\mathbf{Y}$ and $\mathbf{Z}$, respectively. This figure indeed depicts the desired result that all estimated probability intervals have width of one half.

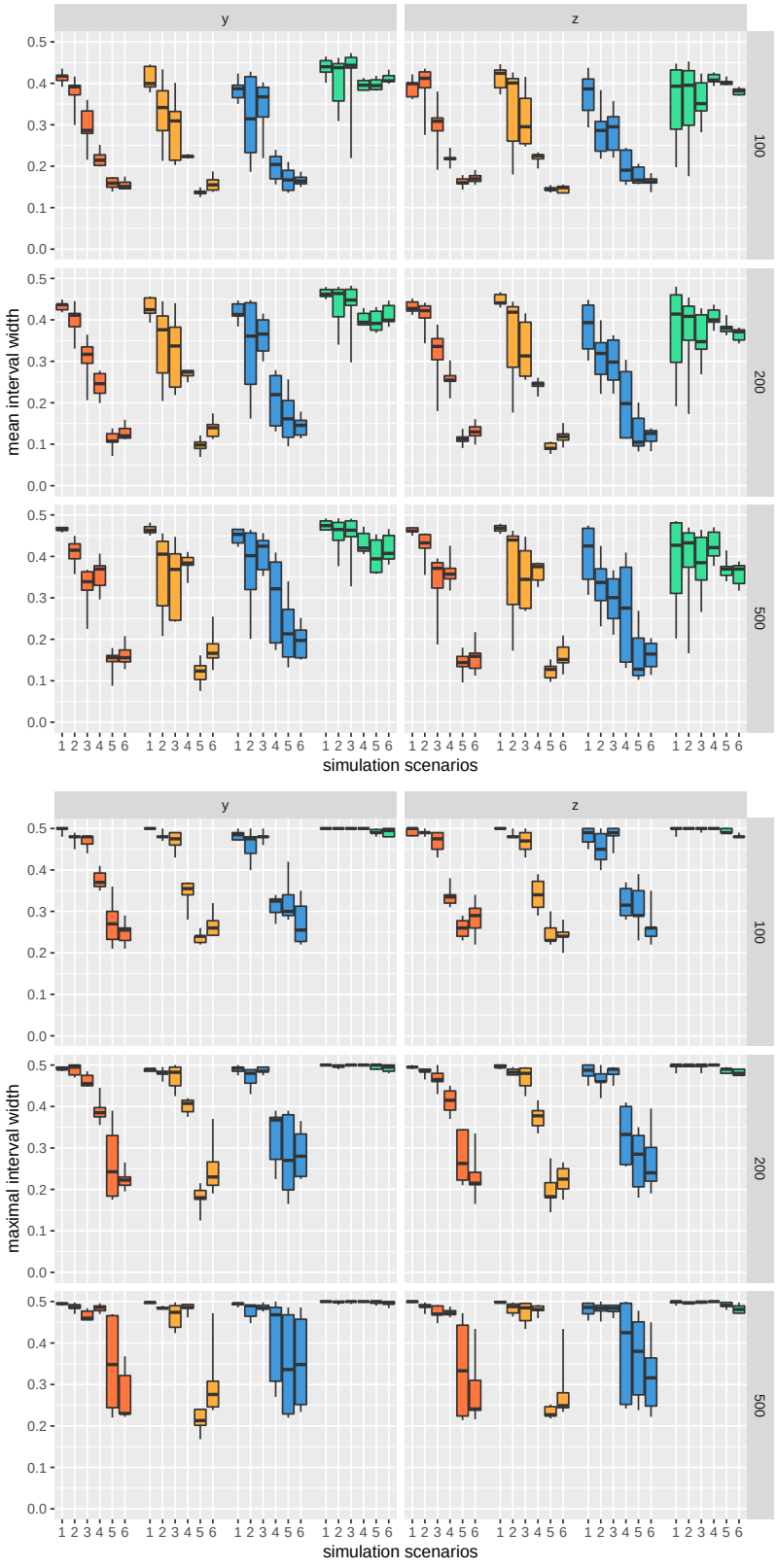

Figure 4: Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables $\mathbf{Y}$ and $\mathbf{Z}$, respectively.

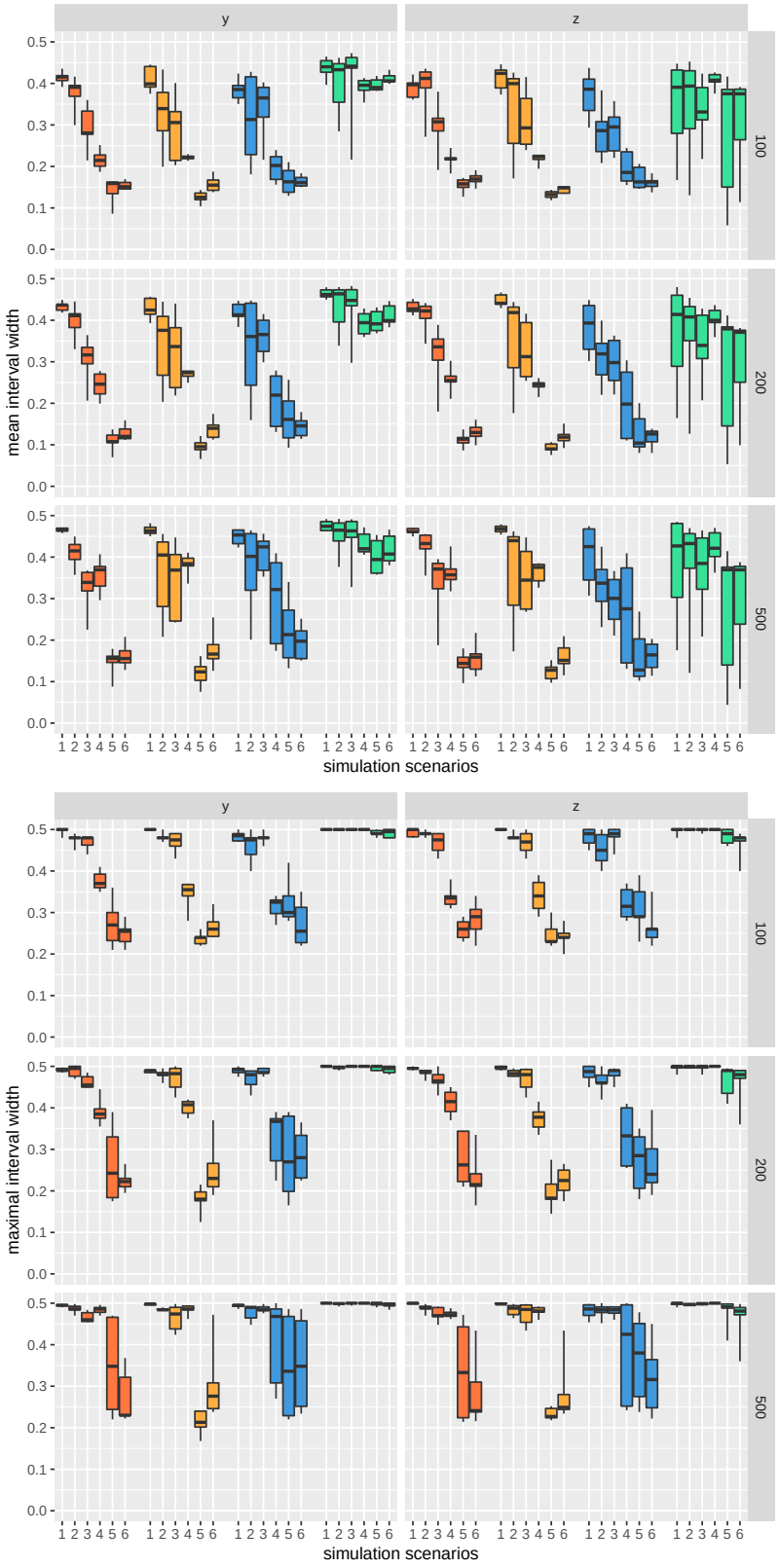

Figure 5: Mean and maximal interval widths of the components of the marginal distributions of the specific variables for case-wise imputation. The two columns display the pooled results for the marginals of the specific variables $\mathbf{Y}$ and $\mathbf{Z}$, respectively.

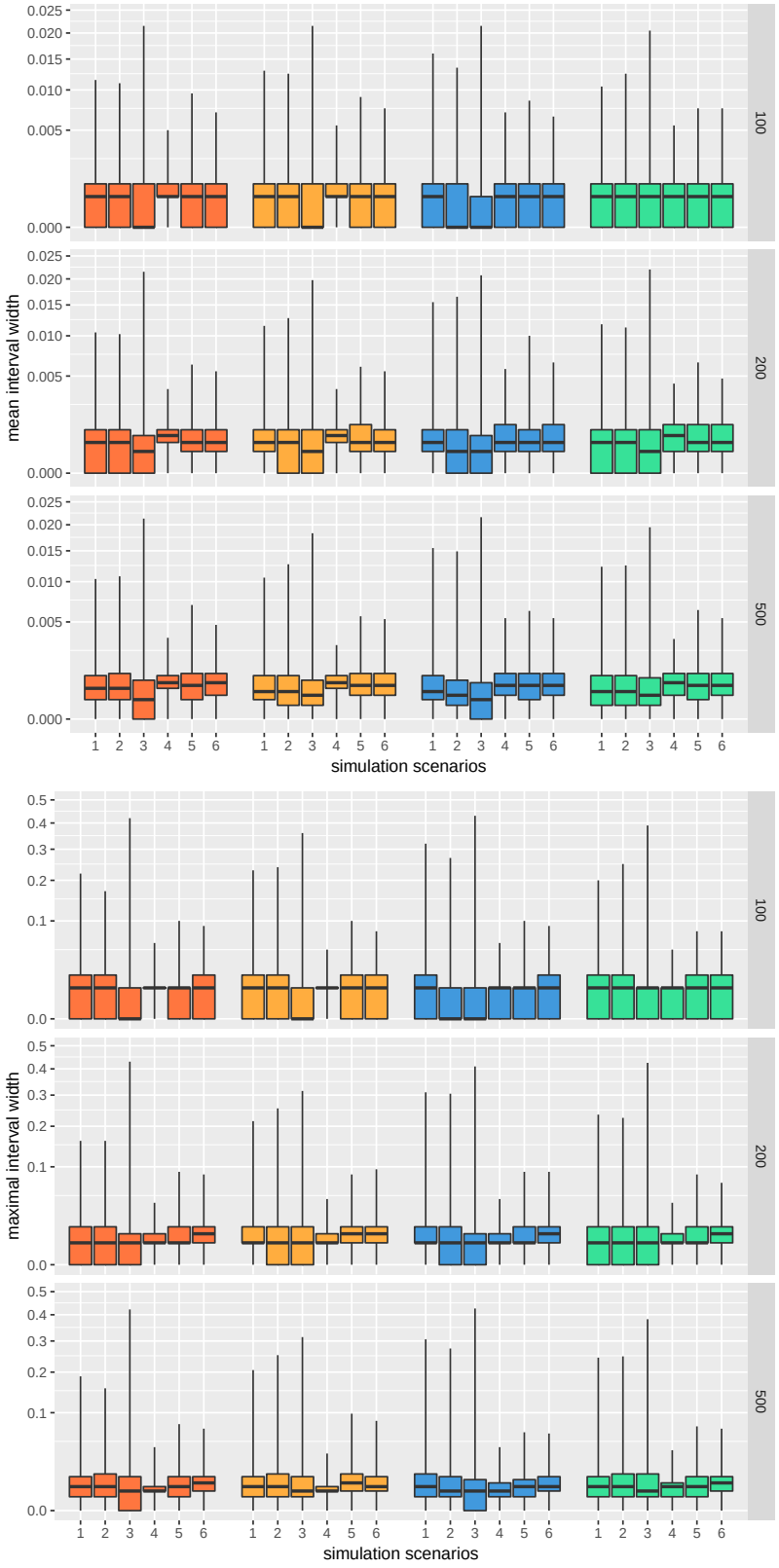

Figure 6: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$) for domain imputation.

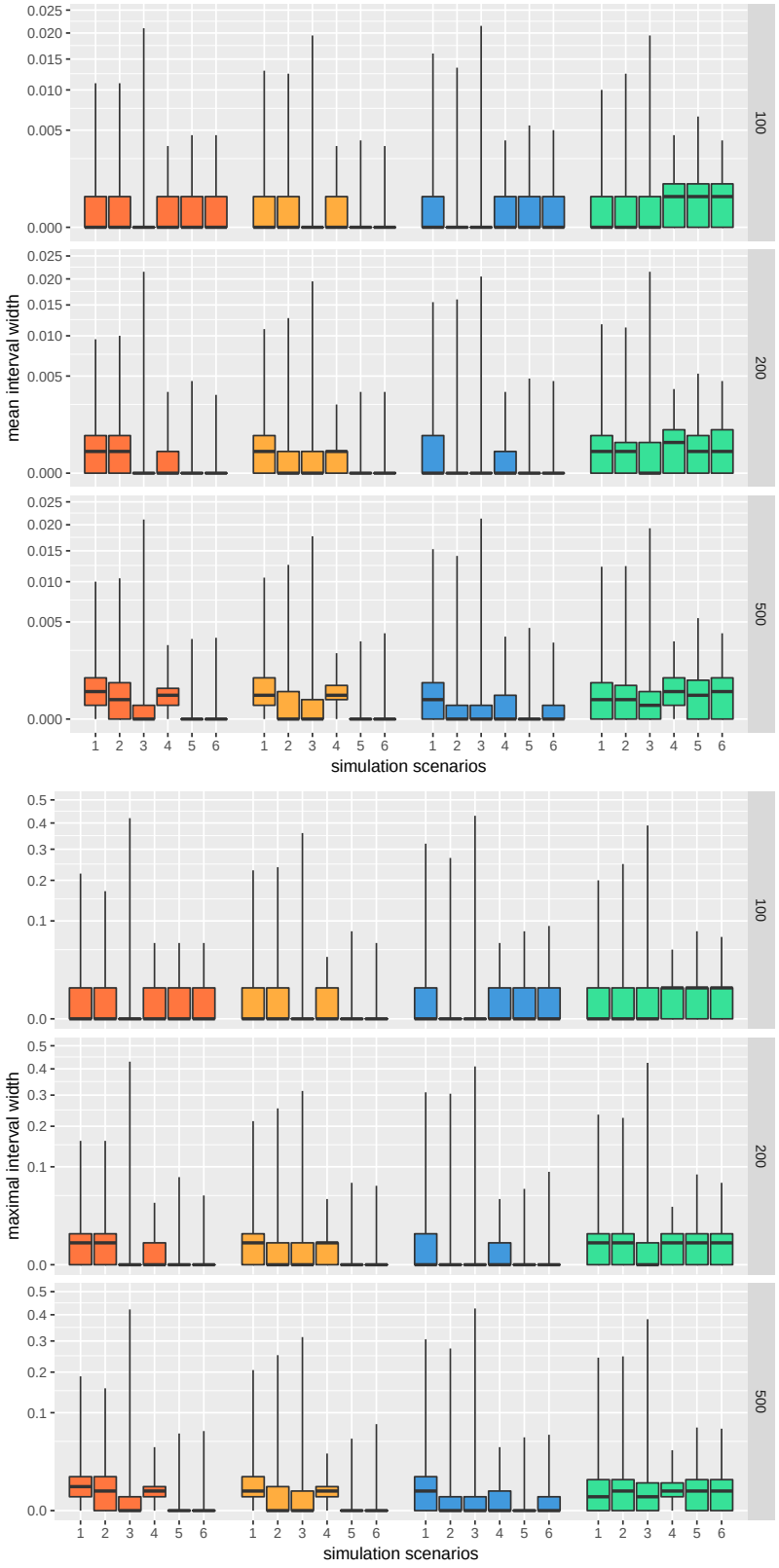

Figure 7: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$) for variable-wise imputation.

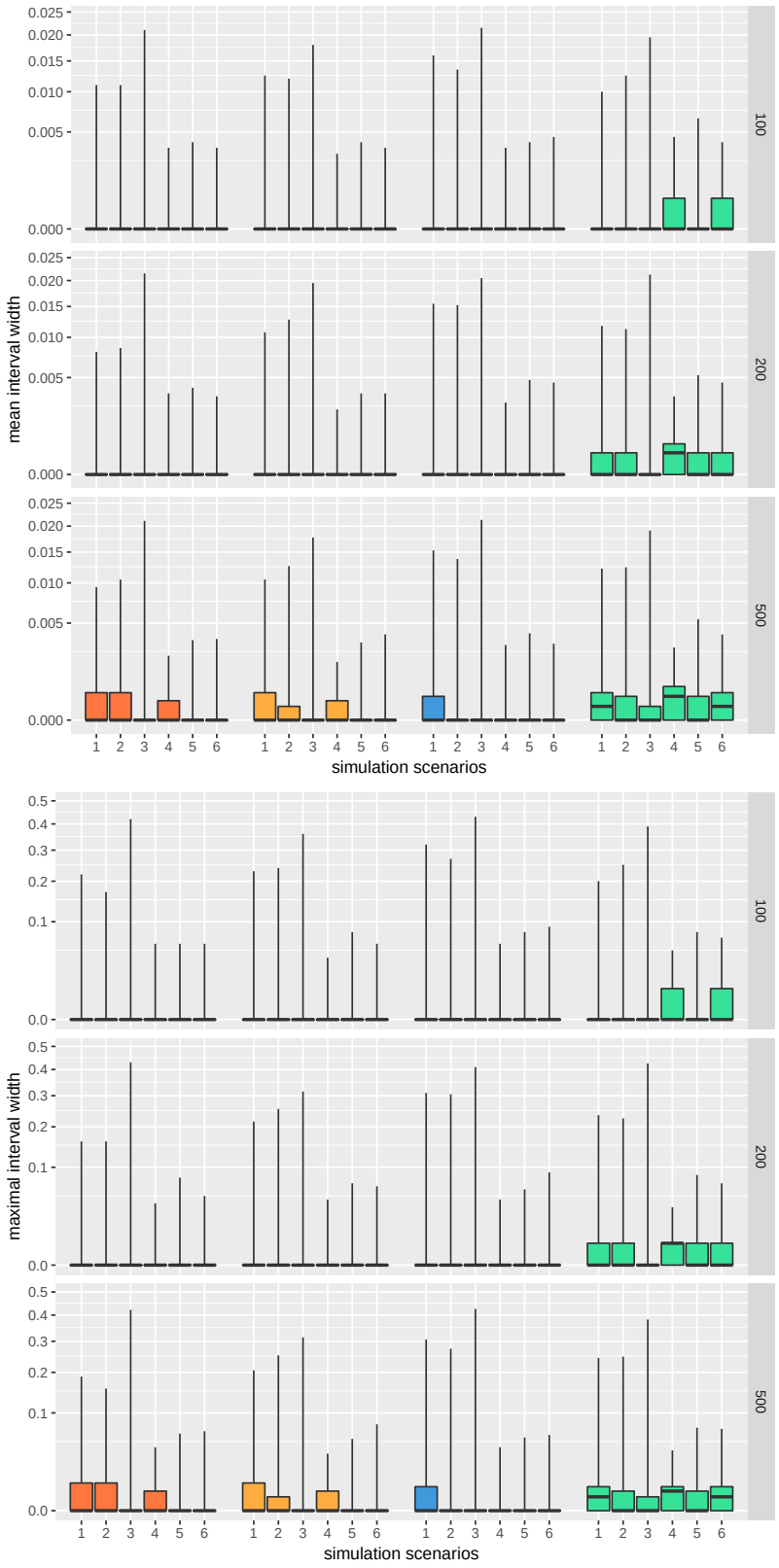

Figure 8: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of (**X**, **Y**, **Z**) for case-wise imputation.