IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2018, accepted December 31, 2018, date of publication January 8, 2019, date of current version February 4, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2891360*

# Generating Synthetic Missing Data: A Review by Missing Mechanism

**MIRIAM SEOANE SANTOS[1], RICARDO CARDOSO PEREIRA[1], ADRIANA FONSECA COSTA[1], JASTIN POMPEU SOARES[1], JOÃO SANTOS[2,3], AND PEDRO HENRIQUES ABREU[1]**

[1]Centre for Informatics and Systems, Department of Informatics Engineering, University of Coimbra, 3030-790 Coimbra, Portugal
[2]Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Center (CI-IPOP), 4200-072 Porto, Portugal
[3]Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, 4050-313 Porto, Portugal

Corresponding author: Miriam Seoane Santos (miriams@student.dei.uc.pt)

This work was supported in part by the Project (Norte Portugal Regional Operational Programme—NORTE 2020) under Grant NORTE-01-0145-FEDER-000027 and in part by the FCT Research Grant under Grant SFRH/BD/138749/2018.

**ABSTRACT** The performance evaluation of imputation algorithms often involves the generation of missing values. Missing values can be inserted in only one feature (univariate configuration) or in several features (multivariate configuration) at different percentages (missing rates) and according to distinct missing mechanisms, namely, missing completely at random, missing at random, and missing not at random. Since the missing data generation process defines the basis for the imputation experiments (configuration, missing rate, and missing mechanism), it is essential that it is appropriately applied; otherwise, conclusions derived from ill-defined setups may be invalid. The goal of this paper is to review the different approaches to synthetic missing data generation found in the literature and discuss their practical details, elaborating on their strengths and weaknesses. Our analysis revealed that creating missing at random and missing not at random scenarios in datasets comprising qualitative features is the most challenging issue in the related work and, therefore, should be the focus of future work in the field.

**INDEX TERMS** Data preprocessing, missing data, missing data generation, missing data mechanisms.

## I. INTRODUCTION

Missing Data (MD) consists of the existence of absent observations (values) in data and is a common obstacle researchers face in real-world contexts [1]–[3]. MD occurs in a variety of domains, for several different reasons, and regardless of whatever they might be, has serious implications for knowledge extraction and classification performance. When datasets are incomplete, pattern classification turns into a more complex task; therefore, over the years, researchers have invested in developing effective strategies to replace the missing values by plausible substitute values, a process generally designated by *data imputation* [4].

A classical approach to data imputation studies follows 4 mains steps (Figure 1):

1) Collection of several complete datasets to perform the experiments. Depending on the nature of the domain, these datasets may encompass several feature types (e.g. qualitative/quantitative) and different dimensionality (number of features and number of patterns);
2) Synthetic generation of missing data. Missing values can be generated in only one feature (univariate configuration) or several features (multivariate configuration), at several percentages (missing rates). Furthermore, the generation may follow 3 different underlying mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) [5];
3) Data imputation using several strategies: common choices rely on *statistical-based* methods (e.g. mean/mode imputation) or *machine learning-based* methods (e.g. KNN imputation) [6];
4) Evaluation of imputation algorithms, either in terms of classification performance (e.g. AUC values) [2] or quality of imputation (e.g. RMSE values) [7], by comparing the substitute values with the ground truth (known original values).

This review focuses on Step 2 – Missing Data Generation – by discussing the existing approaches found in the literature. Over the years, a great effort has been done in what concerns the comparison of different approaches to handle MD (deletion, imputation, model-based approaches) [6], [8], [9]), with a special emphasis on the evaluation of new machine learning methods for imputation (Steps 3 and 4) [10]. However, the process of missing data generation strongly conditions the validity of the conclusions derived from the following steps. If the MD generation approach is ill-defined, some

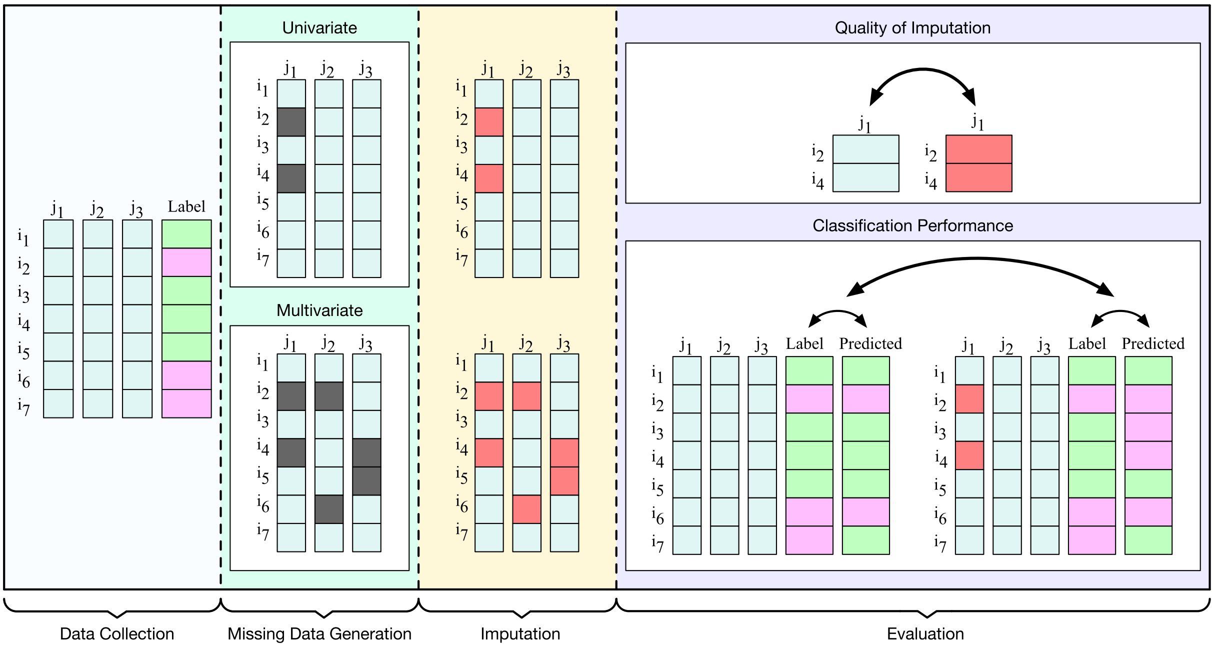


**FIGURE 1.** **Classical experimental setup in data imputation studies.**

hitches may arise during the experimental setup (e.g. the desired missing rate may not be achieved for some scenarios, the mechanisms under which data should be missing may be broken). Thus, the established missing data setup may deviate from what was intended by the researcher, causing the derived conclusions to be biased or invalid. In sum, although the evaluation of different methods to synthetically generate MD remains an understudied topic, it is of crucial importance since they define the working ground for the missing data experiments. The goal of this paper is to illustrate several approaches to missing data generation, thoroughly analyse their practical details and discuss their application in real-world contexts from a theoretical and empirical perspective. To the extent of the authors' knowledge, there is no systematic research on the assessment and evaluation of missing data generation approaches, which constitutes the novelty of this work. The contributions of this research are as follows:

- Providing a thorough analysis of the practical details of each approach and uncovering some issues that may arise during their application;
- Discussing the limitations and restrictions of each approach (e.g. maximum possible MR that they are able to generate);
- Explaining the MR assumptions of each approach (i.e., whether MR is defined for the entire dataset or for a single feature) and presenting the necessary MR adjustments accordingly;
- Suggesting some modifications to the original approaches and elaborating on some implementation details left undiscussed in the original papers.

Considering the contributions given above, this review could prove instrumental for researchers from the Machine Learning field as well as for researchers farther from this field. Researchers familiarized with the missing data topic may learn from an extensive analysis on missing data generation algorithms (their benefits, flaws and limitations) while researchers farther from this topic encounter a complete review where the key concepts on missing data theory, as well as several approaches to missing data generation, are well described and illustrated recurring to schemas and practical examples.

The structure of the paper is as follows: Section II starts by introducing some important notation that will be used throughout the paper, whereas Section III formally describes and illustrates the existing missing data mechanisms. Then, in Sections IV and V, we review several univariate and multivariate implementations for missing data generation that are generic and applicable in several domains, and thoroughly analyse and compare them (by missing mechanism and configuration) in Section VI. Section VII discusses some domain-specific missing data generation approaches, tailored to the peculiarities of a given context, while Section VIII summarizes the key issues one might face when performing experiments using the reviewed generic approaches and discusses the advantages/disadvantages of domain-specific approaches. Finally, Section IX concludes the paper and outlines some potential directions for future research.

## II. PRELIMINARY NOTATION

In order to provide a formal description of the missing data mechanisms, it is first necessary to establish some basic notation and terminology. Let us assume a dataset $\mathbf{X}$ represented by a $n \times p$ matrix, where $i = 1, \cdots, n$ patterns and $j = 1, \cdots, p$ features. The elements of $\mathbf{X}$ are denoted by $x_{i,j}$, each individual feature in $\mathbf{X}$ is denoted by $x_j$ and each pattern is referred to as $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,j}, \cdots, x_{i,p}]$. In classification and missing theory domains, each pattern is also assigned a target class $t_i \in \{C_1, C_2, \cdots, C_c\}$ and a missing indicator $\mathbf{m}_i = [m_{i,1}, m_{i,2}, \cdots, m_{i,j}, \cdots, m_{i,p}]$, which indicates the features that are missing for each pattern $\mathbf{x}_i$. We can now define a missing data indicator $\mathbf{M}$ as a $n \times p$ binary matrix, defined as follows [11]:

$$\mathbf{M} = \{m_{i,j}\}_{i,j=1}^{n,p} = \begin{cases} m_{i,j} = 1, \text{ if } x_{i,j} \text{ is missing} \\ m_{i,j} = 0, \text{ if } x_{i,j} \text{ is observed} \end{cases} \quad (1)$$

$\mathbf{M}$ indicates the locations of the missing values in the dataset and $\mathbf{X}$ may be divided into two components, $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$. $\mathbf{X}_{obs}$ and $\mathbf{X}_{miss}$ represent, respectively, the observed and missing values in $\mathbf{X}$, i.e., $\mathbf{X}_{obs}$ contains all elements $x_{i,j}$ where $m_{i,j} = 0$ while $\mathbf{X}_{miss}$ contains all elements $x_{i,j}$ where $m_{i,j} = 1$. Rubin's missing data theory [12], [13] establishes that the probability distribution of $\mathbf{M}$ may depend on $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$, and that this relationship describes the missing data mechanisms, $p(\mathbf{M} \mid \mathbf{X}, \xi)$, whose parameters are herein denoted by $\xi$ [14], [15]. In practice, $\xi$ cannot be determined with certainty; however, it is not important to know these parameters in detail, it is only necessary to understand whether there is or there is not a relation between $\mathbf{M}$ and $\mathbf{X}$ components: $\mathbf{X}_{obs}$ and $\mathbf{X}_{miss}$.

A dataset $\mathbf{X}$ can suffer from different percentages of missing data, which are referred to as *missing rates* (MRs) and they can be defined for each feature individually or for the entire dataset. Consider Table 1, which illustrates the concepts presented above. Table 1a represents the matrix of data $\mathbf{X}$, where the number of patterns is $n = 20$ (20 records/lines in the table), and the number of features is $p = 2$ ("Age" and "Number of Cigarettes"). Only feature $x_2$ ("Number of Cigarettes") has missing values, denoted by "$\otimes$", but there are several patterns that contain missing values, $\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_8, \mathbf{x}_{10}, \mathbf{x}_{13}, \mathbf{x}_{15}, \mathbf{x}_{17}, \mathbf{x}_{18}\}$. Table 1b represents the missing data indicator matrix $\mathbf{M}$, where positions $x_{ij}$ of Table 1a are coded as 0/1 values according to their presence/absence. As an example, $M_{2,1} = 0$ since "Age" is observed in pattern $\mathbf{x}_2$, while $M_{2,2} = 1$ since "Number of Cigarettes" is missing in $\mathbf{x}_2$. Regarding the MR, feature $x_1$ has a MR of 0% (there are no missing values in "Age"), and feature $x_2$ has a MR of 40% (out of 20 values, 8 are missing in "Number of Cigarettes", $\frac{8}{20} = 40\%$). We may also define the MR considering the entire dataset, that is, the total of $x_{i,j}$ elements that are missing. In this case, there are a total of *patterns* $\times$ *features* elements ($20 \times 2 = 40$ elements), and 8 of them are missing, thus giving a MR of $\frac{8}{40} = 20\%$, if the entire dataset is considered.

## III. MISSING DATA MECHANISMS

We now formally characterize the different missing data mechanisms, $p(\mathbf{M} \mid \mathbf{X}, \xi)$ [16], illustrating each one with an example. For this purpose, consider Table 2 which represents a simulated dataset of a study regarding adolescent tobacco use, with 20 participants. Feature "Age" is completely observed while the "Number of Cigarettes", is missing according to different mechanisms, as explained in what follows.

**TABLE 1.** Adolescent Tobacco Study: (a) matrix of data X, (b) response indicator matrix M.

(a)

| Age | Number of cigarettes |
|---|---|
| 15 | 2 |
| 15 | ⊗ |
| 15 | ⊗ |
| 16 | 2 |
| 16 | 2 |
| 16 | 4 |
| 16 | 3 |
| 17 | ⊗ |
| 17 | 6 |
| 17 | ⊗ |
| 17 | 5 |
| 17 | 5 |
| 18 | ⊗ |
| 18 | 6 |
| 18 | ⊗ |
| 19 | 3 |
| 19 | ⊗ |
| 19 | ⊗ |
| 20 | 9 |
| 20 | 2 |

(b)

| Age | Number of cigarettes |
|---|---|
| 0 | 0 |
| 0 | 1 |
| 0 | 1 |
| 0 | 0 |
| 0 | 0 |
| 0 | 0 |
| 0 | 0 |
| 0 | 1 |
| 0 | 0 |
| 0 | 1 |
| 0 | 0 |
| 0 | 0 |
| 0 | 1 |
| 0 | 0 |
| 0 | 1 |
| 0 | 0 |
| 0 | 1 |
| 0 | 1 |
| 0 | 0 |
| 0 | 0 |

**TABLE 2.** Missing mechanisms example: a simulated dataset of a study in adolescent tobacco use, where the daily average of smoked cigarettes is missing under different mechanisms (MCAR, MAR, and MNAR).

| Age | Number of cigarettes | | | |
|---|---|---|---|---|
| | Complete | MCAR | MAR | MNAR |
| 15 | 2 | 2 | ⊗ | 2 |
| 15 | 9 | ⊗ | ⊗ | ⊗ |
| 15 | 4 | ⊗ | ⊗ | 4 |
| 16 | 2 | 2 | ⊗ | 2 |
| 16 | 2 | 2 | ⊗ | 2 |
| 16 | 7 | 4 | ⊗ | ⊗ |
| 16 | 3 | 3 | ⊗ | 3 |
| 17 | 9 | ⊗ | 9 | ⊗ |
| 17 | 6 | 6 | 6 | ⊗ |
| 17 | 4 | ⊗ | 4 | 4 |
| 17 | 5 | 5 | 5 | 5 |
| 17 | 5 | 5 | 5 | 5 |
| 18 | 7 | ⊗ | 7 | ⊗ |
| 18 | 6 | 6 | 6 | ⊗ |
| 18 | 7 | ⊗ | 7 | ⊗ |
| 19 | 3 | 3 | 3 | 3 |
| 19 | 8 | ⊗ | 8 | ⊗ |
| 19 | 3 | ⊗ | 3 | 3 |
| 20 | 9 | 9 | 9 | ⊗ |
| 20 | 2 | 2 | 2 | 2 |

In Missing Completely At Random (MCAR) mechanism, $\mathbf{M}$ is completely unrelated to the input data $\mathbf{X}$ – completely unrelated to both $\mathbf{X}_{obs}$ and $\mathbf{X}_{miss}$ (2). For MCAR, the probability of missingness depends only on parameters $\xi$; or in other words, the probability of missing values in a feature $x_j$ is completely random. Considering Table 2, MCAR values were produced by random deletion: the missing values are not located in a particular range of "Age" or "Number of Cigarettes" values. This mechanism can, therefore, be due to unexpected events during the study: a participant had a flat tire and could not attend the appointment or was with the flue.

$$p(\mathbf{M} = 1 \mid \mathbf{X}, \xi) = p(\mathbf{M} = 1 \mid \xi) \quad (2)$$

Missing At Random (MAR) mechanism occurs when the probability of missingness depends on the observed information $\mathbf{X}_{obs}$, but not on $\mathbf{X}_{miss}$ (3). In other words, the probability of missing values in a feature $x_j$ may depend on the observed values of other features in the dataset, but not on the values of $x_j$ itself. In Table 2, MAR scenario is created by the missing values of "Number of Cigarettes" for younger participants (aged between 15 and 16 years). It could be the case that younger adolescents are less likely to fill in their number of smoked cigarettes per day because they do not want to admit that they are regular smokers. However, the missingness is unrelated to the number of cigarettes smoked by these teenagers, had it been reported (note the "Complete" column, where a low and high number of cigarettes would be found among the missing values, had they been observed). The probability of missing values in "Number of Cigarettes" is therefore a function of the observed information $\mathbf{X}_{obs}$ only, unrelated to the missing values in the study, $\mathbf{X}_{miss}$.

$$p(\mathbf{M} = 1 \mid \mathbf{X}, \xi) = p(\mathbf{M} = 1 \mid \mathbf{X}_{obs}, \xi) \tag{3}$$

Finally, in Missing Not At Random (MNAR) mechanism, the missingness may depend on both observed and unobserved information – both $\mathbf{X}_{obs}$ and $\mathbf{X}_{miss}$ – and the general expression of the missing data model cannot be simplified (4). In a simple manner, this means that the probability of missing values occurring in a feature $x_j$ may be related to the observed values of other features in the dataset ($\mathbf{X}_{obs}$), as well as the underlying, unknown values of $x_j$ itself ($\mathbf{X}_{miss}$). In Table 2, MNAR values are missing for higher values of "Number of Cigarettes": the probability of missing values in "Number of Cigarettes" is related to the missing values themselves, had they been observed (note the "Complete" column). This would be the case of teenagers that refused to report their number of smoked cigarettes per day since they smoked a very large quantity.

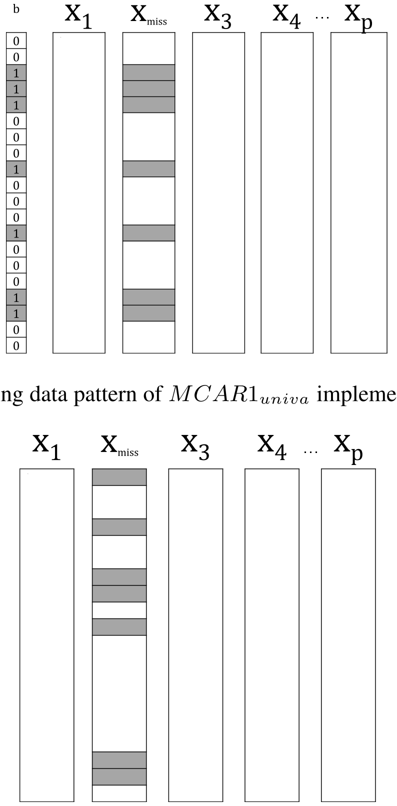
$$p(\mathbf{M} = 1 \mid \mathbf{X}, \xi) = p(\mathbf{M} = 1 \mid \mathbf{X}_{obs}, \mathbf{X}_{miss}, \xi) \tag{4}$$

## IV. UNIVARIATE CONFIGURATIONS

Univariate configurations, herein designated by *univa* configurations, refer to those where only one feature in the study suffers from missing data. These *univa* configurations contrast with the *unifo* configurations (explained in the next section), where the missing values affect several (if not all) features in the dataset. The terms *univa* and *unifo* were taken from the research of Twala [17], one of the first works regarding the synthetisation of missing data mechanisms. We therefore begin this section with the *univa* implementations of MCAR, starting with the algorithm proposed by Twala [17].

### A. UNIVARIATE MCAR IMPLEMENTATIONS

The MCAR *univa* implementation of Twala [17] considers that the feature to be missing, $x_{miss}$, should be the one most correlated with the class labels $t$. Furthermore, Twala et al. considered the definition of MR as the percentage of missing values over the entire dataset, as explained in Section II.



(a) Missing data pattern of $MCAR1_{univa}$ implementation.

(b) Missing data pattern of $MCAR2_{univa}$ implementation.

**FIGURE 2. Schemes describing missing data patterns of each MCAR implementation. The shaded observations represent the location of missing values in the dataset. In (a), the randomness is defined by the Bernoulli distribution, represented by vector *b*.**

To respect the overall MR specified, the individual percentage of missing values in the chosen feature must be adjusted: for an overall percentage of MR% (over the entire dataset), an individual feature must have $p$ × MR% of missing values, with $p$ being the number of features in $\mathbf{X}$.

To determine which elements should be missing in $x_{miss}$, a Bernoulli distribution is used. The Bernoulli distribution is a discrete distribution that has outcome 1 with probability *prob* and outcome 0 with probability $1 - prob$, as shown in (5). The missing elements of $x_{miss}$ are chosen by performing $n$ Bernoulli trials with probability of success *prob*, with $n$ being the number of patterns in the dataset and *prob* being the expected MR. Thus being, each pattern is associated with a probability of success (probability of being missing) equal to MR (Figure 2a).

$$f(k, prob) = \begin{cases} 1 - prob \text{ for } k = 0 \\ prob \text{ for } k = 1 \end{cases} \tag{5}$$

A different MCAR *univa* implementation was proposed by Rieger *et al.* [18] and Xia *et al.* [19], where random locations of $x_{miss}$ are chosen (using a random number generator) and their values are deleted (Figure 2b).

Finally, García-Laencina *et al.* [6], [10] consider a MCAR *univa* implementation where $x_{miss}$ is either chosen randomly

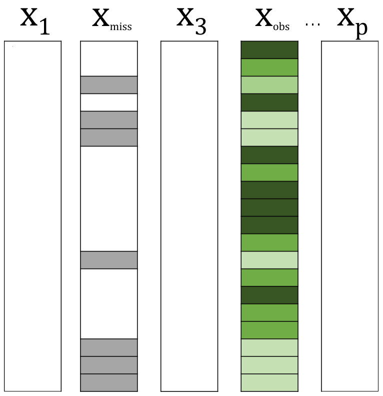


**FIGURE 3.** **Missing data pattern of $MAR1_{univa}$ implementation. Shaded observations represent the location of missing values in $x_{miss}$, whereas the magnitude of $x_{obs}$ values is represented by different shades of green, with dark green indicating higher values and light green indicating lower values. In $MAR1_{univa}$, values of $x_{miss}$ are missing for lower values of $x_{obs}$.**

or according to its relevance for classification. In this implementation, the "relevance" of a feature is determined by the Normalized Mutual Information (NMI) between such feature and the classification target [10]. The missing values are randomly introduced in the feature of interest, $x_{miss}$, and the missing rate is specified for that feature only (MR% of missing values in that feature alone, not in the entire dataset).

### B. UNIVARIATE MAR IMPLEMENTATIONS

Regarding MAR *univa*, five different implementations are reviewed – $MAR1_{univa}$, $MAR2_{univa}$, $MAR3_{univa}$, $MAR4_{univa}$ and $MAR5_{univa}$ – following the research works of Twala [17], Rieger *et al.* [18], and Xia *et al.* [19]. All MAR implementations make use of an observed, *determining* feature, $x_d$ or $x_{obs}$ (also referred to as a *causative* feature in some works [20]), which defines the missing locations in $x_{miss}$. An example is given in Figure 3, where the missing positions in $x_{miss}$ are influenced by the corresponding values of $x_{obs}$.

$MAR1_{univa}$ refers to the research work of Twala [17], and similar to the MCAR *univa* implementation, the feature most correlated with the class labels is chosen as $x_{miss}$. Then, among the remaining features, the one most correlated with $x_{miss}$ is chosen to be the determining feature $x_{obs}$. As explained for $MCAR1_{univa}$, the individual feature $x_{miss}$ must have $p \times$ MR% of missing values, since in the implementations suggested by Twala et al. the MR is defined for the entire dataset.

After the pair of correlated features $\{x_{miss}, x_{obs}\}$ is found, the locations where $x_{miss}$ will be missing are then defined according to the values of $x_{obs}$. Let us define a variable $k$ that represents the necessary MR adjustment, $k = p \times \text{MR}$. The value of $k$% will define the percentile of $x_{obs}$ that must be found in order to produce the missing values in $x_{miss}$: values of $x_{miss}$ lower than the $k$% percentile of $x_{obs}$ are set to be missing. In other words, the percentile of $k$% returns the cut-off value for which $k$% of $x_{obs}$ are lower than that cut-off. As an example, consider an overall MR of 45% and the pair of features $\{x_1, x_2\}$, where $x_{obs}$ is $x_1$ and $x_{miss}$ is $x_2$. The missing locations in $x_2$ will be determined by the $p \times$ MR% $=$ 90% percentile of $x_1$. Imagine that the 90% percentile of $x_1$ is 3.4: values of $x_2$ where the corresponding values $x_1$ are lower than 3.4 will be set to missing values. Thus being, $x_2$ will have a total of 90% of missing values, resulting in an overall (0+90)/2 $=$ 45% MR, as specified. Figure 3 shows a pictorial example of $MAR1_{univa}$ where the light green positions represent the lowest values of $x_{obs}$, where $x_{miss}$ is missing.

Rieger *et al.* [18] propose implementations $MAR2_{univa}$ to $MAR5_{univa}$. $MAR2_{univa}$ is based on the ranks of $x_{obs}$ ($r_{obs}$): the probability of an element $x_{i,miss}$ to be missing is computed by dividing the rank of $x_{i,miss}$ in the determining feature $x_{obs}$ by the sum of all ranks of $x_{obs}$ (6). This is also the implementation proposed by Xia *et al.* [19].

$$P(x_{i,miss} = \text{missing}) = \frac{r_{i,obs}}{\sum_{i=1}^{n} r_{i,obs}} \tag{6}$$

The patterns to have missing values in $x_{miss}$ are then sampled according to their resulting probability $P(x_{i,miss})$. The choice of $x_{miss}$ and $x_{obs}$ is arbitrary and can either be random or specified by the researcher. Furthermore, the definition of MR is not described in the original paper and one might consider a MR for the entire dataset or for each feature individually.

In $MAR3_{univa}$, the patterns are divided into two groups according to the median of the determining feature $x_{obs}$, so that the probability of missingness is different among groups according to (7) ($nG_1$ and $nG_2$ are the number of patterns in Group 1 and Group 2, respectively). Again, the patterns are sampled according to the established probability of missingness (8).

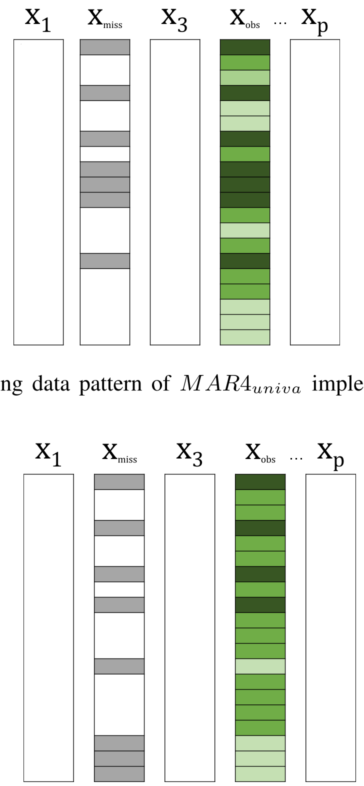
$$\begin{cases} \text{if } x_{i,obs} \geq median(x_{obs}), \text{ then } x_{i,obs} \in G_1 \\ \text{if } x_{i,obs} < median(x_{obs}), \text{ then } x_{i,obs} \in G_2 \end{cases} \tag{7}$$

$$\begin{cases} \text{if } x_{i,obs} \in G_1 \implies P\left(x_{i,miss} = \text{missing}\right) = \dfrac{0.9}{nG_1} \\ \text{if } x_{i,obs} \in G_2 \implies P\left(x_{i,miss} = \text{missing}\right) = \dfrac{0.1}{nG_2} \end{cases} \tag{8}$$

In $MAR4_{univa}$, the locations of $x_{miss}$ that will be missing are chosen according to the positions where $x_{obs}$ assumes its highest values (Figure 4a). $MAR5_{univa}$ considers both the highest and lowest values of $x_{obs}$: given the necessary number of elements to have missing values for the specified MR, call it $N$, $MAR5_{univa}$ sets $N/2$ elements to have missing values according to the highest values of $x_{obs}$, and $N/2$ according to the lowest (Figure 4b).

### C. UNIVARIATE MNAR IMPLEMENTATIONS

For MNAR mechanism, we refer to the implementations of Twala [17] ($MNAR1_{univa}$) and Xia *et al.* [19] ($MNAR2_{univa}$). These approaches are similar: in $MNAR1_{univa}$, the lowest values of $x_{miss}$ are set to be missing, until the desired MR is achieved; in $MNAR2_{univa}$, the same procedure is applied, although the highest values are considered instead.



(a) Missing data pattern of $MAR4_{univa}$ implementation.

(b) Missing data pattern of $MAR5_{univa}$ implementation.

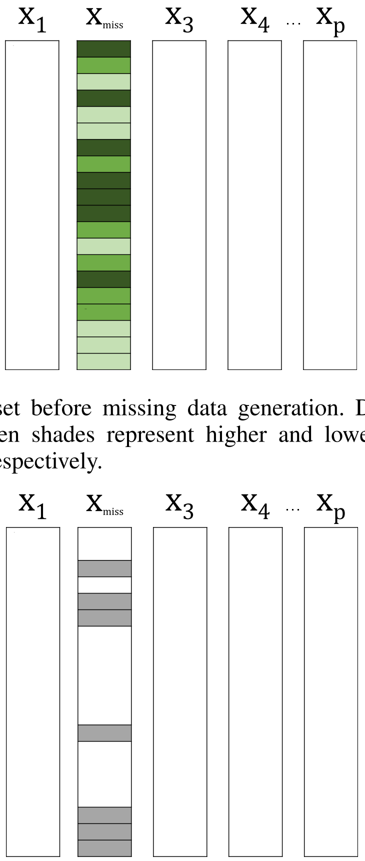
**FIGURE 4. Schemes describing missing data patterns of each MAR implementation. The shaded observations represent the location of missing values in the missing feature. For the observed feature, the values are represented with different shades of green: darker shades are used to represent higher values while lighter shades represent lower values. (a) Missing data pattern of $MAR4_{univa}$ implementation. (b) Missing data pattern of $MAR5_{univa}$ implementation.**

$MNAR1_{univa}$ is illustrated in Figure 5, where missing locations of $x_{miss}$ (Figure 5a) are conditioned by the values of $x_{miss}$ itself (Figure 5b): missing values are inserted where $x_{miss}$ assumes lower values (light green). In $MNAR2_{univa}$, the highest $x_{miss}$ values are deleted until the desired MR is achieved (Figure 6).

Similar to the above-mentioned approaches by Twala [17], $x_{miss}$ is the feature most correlated with the class labels. Then, $x_{miss}$ itself is used as a determining feature; the $k$% percentile of $x_{miss}$ is determined and values lower than the cut-off value are set to be missing.

## V. MULTIVARIATE CONFIGURATIONS

In multivariate configurations, which we denote by *unifo* configurations, the missing values are generated in all features, with the exception of MAR mechanism. For MAR there are two common approaches, as will be illustrated in Section V-B: i) choosing one determining feature $x_{obs}$ that will define the missing positions in the remaining features or ii) creating pairs of correlated features $\{x_{obs}, x_{miss}\}$ where the missing values in $x_{miss}$ of each pair are defined by the corresponding $x_{obs}$ feature.



(a) Dataset before missing data generation. Dark and light green shades represent higher and lower $x_{miss}$ values, respectively.

(b) Dataset after missing data generation. The shaded observations represent the location of missing values in the missing feature.

**FIGURE 5. Missing data pattern of $MNAR1_{univa}$ implementation. (a) Dataset before missing data generation. Dark and light green shades represent higher and lower $x_{miss}$ values, respectively. (b) Dataset after missing data generation. The shaded observations represent the location of missing values in the missing feature.**

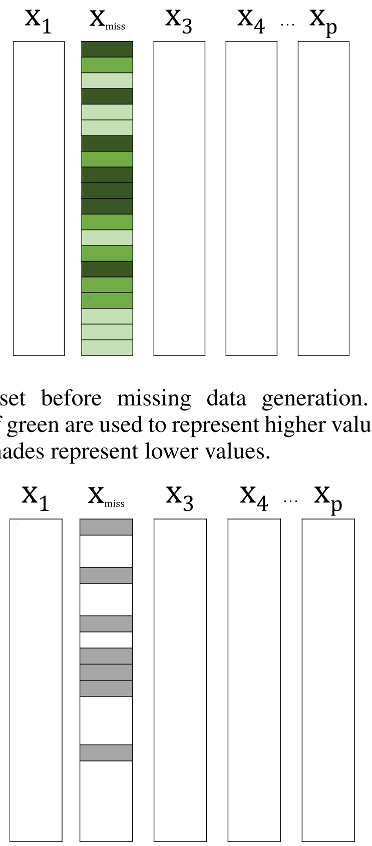
### A. MULTIVARIATE MCAR IMPLEMENTATIONS

$MCAR_{unifo}$ implementations are an extension of $MCAR_{univa}$ implementations, where all elements $x_{i,j}$ are eligible to be deleted, instead of focusing only on a feature $x_{miss}$. Herein, we refer to three $MCAR_{unifo}$ implementations that follow naturally from the *univa* configurations.

We start with $MCAR1_{unifo}$, proposed by Twala [17]. In $MCAR1_{unifo}$, all features will have the same percentage of missing values, specified by MR: $n$ Bernoulli trials are generated for each feature $p$ in the dataset and the missing elements $x_{i,j}$ are determined accordingly. In other words, $x_{i,j}$ is missing if $b_{i,j} = 1$, where $b$ indicates the 1/0 outcome for each trial (Figure 7).

$MCAR2_{unifo}$ follows from the research works of Garciarena and Santana [20], Zhu *et al.* [21], Pan *et al.* [22], and Ali and Omer [23]. In $MCAR2_{unifo}$, a number of $N$ elements $x_{i,j}$ are randomly and deleted (Figure 8).

The MR is defined for the entire dataset and therefore $N = n \times p \times$ MR. However, unlike $MCAR1_{unifo}$, the



(a) Dataset before missing data generation. Darker shades of green are used to represent higher values while lighter shades represent lower values.

(b) Dataset after missing data generation. The shaded observations represent the location of missing values in the missing feature.

**FIGURE 6. Missing data pattern of $MNAR2_{univa}$ implementation. (a) Dataset before missing data generation. Darker shades of green are used to represent higher values while lighter shades represent lower values. (b) Dataset after missing data generation. The shaded observations represent the location of missing values in the missing feature.**

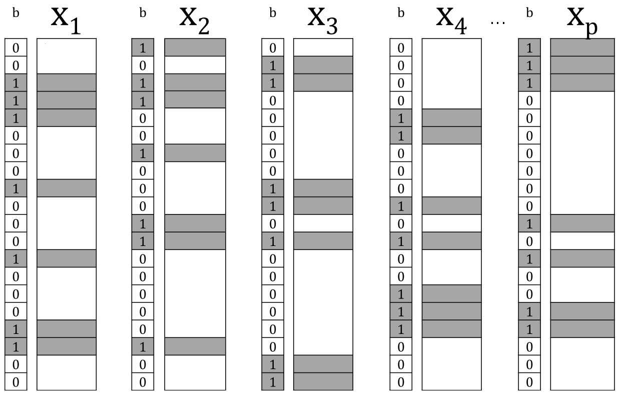


**FIGURE 7. Missing data pattern of $MCAR1_{unifo}$ implementation. *b* represents the Bernoulli distribution for each feature.**

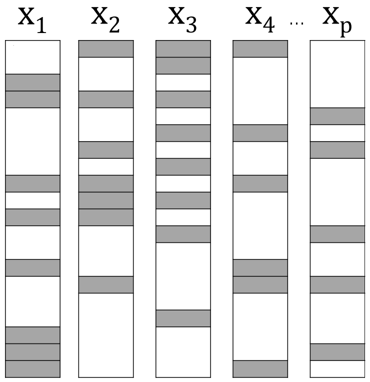


**FIGURE 8. Missing data pattern of $MCAR2_{unifo}$ implementation.**

features are not required to have the same number of missing values, given that all $x_{i,j}$ are eligible for missing data generation and they are chosen randomly across all features. Given the variability of possible missing datasets that can be generated with this approach (more than for $MCAR1_{unifo}$), it is fundamental that missing data experiments using it perform several runs [7], as further discussed in Section VI.

### B. MULTIVARIATE MAR IMPLEMENTATIONS

As stated at the beginning of Section V, there are two main approaches in what concerts $MAR_{unifo}$ implementations:

- Consider a determining feature $x_{obs}$ that will determine the missing pattern of the remaining features ($p - 1$ features or a subset of $n_{x_{miss}}$ features), which is the approach proposed by Garciarena and Santana [20];
- Consider several pairs of features $\{x_{obs}, x_{miss}\}$: for each pair, there is a determining feature $x_{obs}$ that defines the missing pattern of its corresponding $x_{miss}$, which is the approach of Twala [17], Ali and Omer [23], Zhu *et al.* [21], and Pan *et al.* [22].

We start by the simplest $MAR_{unifo}$ approach, the one proposed by Garciarena and Santana [20], which we designate by $MAR1_{unifo}$. $MAR1_{unifo}$ considers the desired MR percentage and number of features $n_{x_{miss}}$ losing their values and starts by randomly choosing the determining feature $x_{obs}$ and the missing features $x_{miss}$. Then, similarly to $MAR1_{univa}$, elements of the $x_{miss}$ features corresponding to lower values of $x_{obs}$ are deleted (Figure 9a). Due to the freedom of choosing a given number of $n_{x_{miss}}$, the missing rates that are possible to generate are restricted by the number of existing features $p$ and chosen features $n_{x_{miss}}$, as will be further explained in Section VI.

We follow to the $MAR_{unifo}$ implementation by Twala [17], $MAR2_{unifo}$. As a natural extension of $MAR1_{univa}$, $MAR2_{unifo}$ considers the creation of several correlated pairs where, for each pair, the most correlated feature is chosen as $x_{miss}$ (Figure 9b). As an example, for $\mathbf{X} = \{x_1, x_2, \cdots, x_8\}$, we could define the pairs $\{x_1, x_2\}$, $\{x_3, x_4\}$, $\{x_5, x_6\}$ and $\{x_7, x_8\}$, assuming that $x_1$ is highly correlated with $x_2$, $x_3$ with $x_4$ and so forth. Although Twala et al. [17] do not specify the procedure for an odd number of features (say, 9 features in the previous example), we assume the creation of triples, where the remaining feature (in the example, $x_9$) is added to the pair that includes its most correlated feature. Following the example, assuming that the feature most correlated with $x_9$ is $x_3$, then the triple $\{x_3, x_4, x_9\}$ is created instead of $\{x_3, x_4\}$.

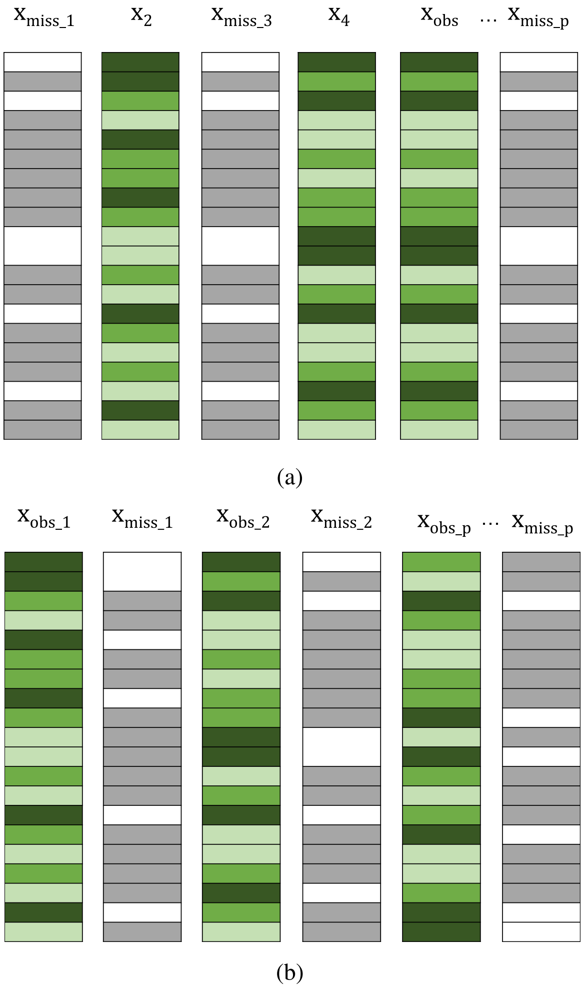


**FIGURE 9. Schemes describing missing data patterns of (a) *MAR*1$_{unifo}$ and (b) *MAR*2$_{unifo}$.**

After the pairs are created, the feature of each pair most correlated with the class labels $t$ is selected to have its values missing; for triples, the two most correlated features with the class labels are chosen.

The desired MR is defined for the entire dataset, but since only one feature will be missing in each pair (or two features in case of triples), the MR must be adjusted for the individual $x_{miss}$ features (9).

$$\text{For an overall MR\%} \begin{cases} k = 2 \times MR\% \text{ for pairs} \\ k = 1.5 \times MR\% \text{ for triples} \end{cases} \tag{9}$$

The positions where each feature $x_{miss}$ will be missing are defined according to the values of $x_{obs}$: for each pair/triple, the k% percentile of $x_{obs}$ is determined. Then, values of $x_{obs}$ lower than the k% percentile are set missing. Similarly to $MAR1_{univa}$, the k% percentile of $x_{obs}$ returns the cut-off value for which k% of $x_{obs}$ are lower than that cut-off. As an example, consider an overall MR of 45% and 5 features already paired: $\{x_1, x_2\}$ and $\{x_3, x_4, x_5\}$, where $x_2$, $x_4$ and $x_5$ are the most correlated with the class labels $t$. The missing positions in $x_2$ will be determined by the 2 × MR = 90% percentile of $x_1$ and the missing positions in $x_4$ and $x_5$ will be determined by the 1.5 × MR = 67.5% percentile of $x_3$. Imagine that the 90% percentile of $x_1$ is 3.4: values of $x_2$ where the corresponding values of $x_1$ are lower than 3.4 will be set missing and $x_2$ individually will have 90% of missing values. The same is performed for $x_4$ and $x_5$. Thus, $x_1$ and $x_3$ will be complete, $x_2$ will have 90% of missing values and $x_4$ and $x_5$ will have 67.5% of missing values each, resulting in an overall $(0+90+0+67.5+67.5)/5 = 45\%$ missing rate.

Ali and Omer [23] propose a similar approach to the above, and we will refer to their approach as $MAR3_{unifo}$. In $MAR3_{unifo}$, the dataset $\mathbf{X}$ is first decomposed into pairs/triples of correlated features, and one feature in each pair/triple conditions the missing pattern of the remaining. However, authors do not elaborate on the choice of which features should be missing and which should be observed; therefore, we assume that the choice may be performed randomly. For each pair/triple, one feature is randomly chosen to be the determining feature $x_{obs}$ and the remaining are therefore the missing features, $x_{miss}$.

Another difference of this approach in comparison to $MAR2_{unifo}$ is that it considers the median of each $x_{obs}$ to define the missing pattern of $x_{miss}$. Given a pair of features $\{x_{obs}, x_{miss}\}$, the median of $x_{obs}$ is determined and two groups are defined: one that contains the positions of $x_{obs}$ whose values are lower than (or equal to) its median and the other containing the positions whose values are higher than its median. Then, one of those groups is randomly selected and will define the missingness of $x_{miss}$ in the following way: given a missing rate MR%, 4× MR% (or 3× MR% for triples) of missing positions are randomly chosen from the group, and the corresponding positions in $x_{miss}$ are set missing.

The $MAR_{unifo}$ approach by Zhu *et al.* [21] and Pan *et al.* [22] ($MAR4_{unifo}$) handles features according to their type. If $x_{obs}$ is continuous or ordinal, the median of $x_{obs}$ is determined and two groups are created, as in the approach by Ali and Omer [23]: one where the values of $x_{obs}$ are lower or equal to the median and other where values of $x_{obs}$ are higher than the median. Otherwise, if $x_{obs}$ is nominal, the existing categories are assigned to two groups of equal size. According to the original paper [21], this assignment is performed by randomly dividing the categories of $x_{obs}$ into two parts, although this does not guarantee that two equally-sized groups are formed, as further detailed in Section VI. After creating the groups, one is randomly chosen and their corresponding values in $x_{miss}$ are set missing with 4 × MR or 3 × MR, for triples.

### C. MULTIVARIATE MNAR IMPLEMENTATIONS

$MNAR_{unifo}$ implementations follow from the $MAR_{unifo}$ implementations discussed in the previous section, proposed in the same research works – Twala [17], Garciarena and Santana [20], Zhu *et al.* [21], Pan *et al.* [22], and Ali and Omer [23]. Similarly, we start by the approach

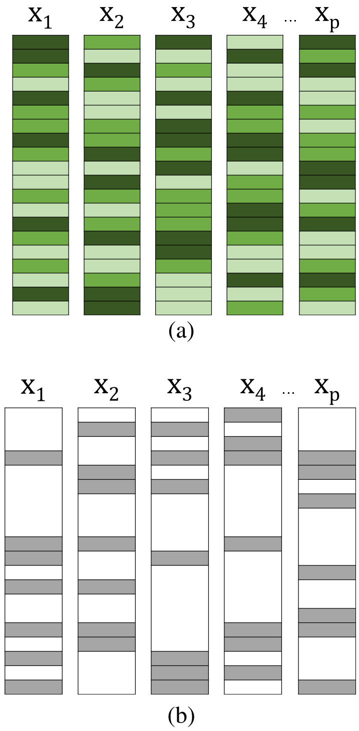


**FIGURE 10. Missing data pattern of $MNAR1_{unifo}$ implementation: (a) Dataset before missing data generation. Darker shades of green represent higher values while lighter shades represent lower values; (b) Dataset after missing data generation. The shaded observations represent the location of missing values in the missing feature.**

presented in Garciarena and Santana [20], herein referred to as $MNAR1_{unifo}$.

Garciarena and Santana [20] propose two MNAR approaches designated MIV and MuOv in the original paper. MIV stands for Missingness depending on its Value Itself and directly illustrates the mechanism explained in Section III, where the probability of a value to be missing depends on the value itself. MuOv (Missing depending on unobserved Variables) is somewhat a domain-based MNAR approach, and therefore we will illustrate it in Section VII. MIV approach (herein designated $MNAR1_{unifo}$) is an extension of $MAR1_{unifo}$, where $x_{obs} = x_{miss}$. In other words, there is not a determining feature $x_{obs}$ that affects the missingness of $x_{miss}$. Instead, the probability of a value to be missing in each feature $x_{miss}$ is determined by the values of each $x_{miss}$ itself. In $MNAR1_{unifo}$, as illustrated in Figure 10, the lowest values of each $x_{miss}$ are found and deleted, according to the specified MR. Similarly to $MAR1_{unifo}$, MR is specified for the entire dataset and the number of features losing their values can be chosen by the researcher.

The $MNAR_{unifo}$ approach proposed by Twala [17], $MNAR2_{unifo}$, follows the same pairing logic as $MAR_{unifo}$. However, the values that are set missing in feature $x_{miss}$ of each pair/triple are defined by the values of $x_{miss}$ itself: lower values of $x_{miss}$ are deleted.

Contrariwise, the $MNAR_{unifo}$ approaches by Ali and Omer [23] ($MNAR3_{unifo}$), Zhu *et al.* [21], and Pan *et al.* [22] ($MNAR4_{unifo}$) do not require the creation of pairs/triples since the missing values are generated directly in all features, according to their respective medians. In $MNAR3_{unifo}$, two groups are defined for each feature, one with values lower or equal to its median and the other with values higher than its median. Then, one group is randomly chosen to have 2 × MR% of missing values, so that the overall MR% over the entire dataset is kept. Similarly, in $MNAR4_{unifo}$, missing values are inserted directly in each feature, without the need of creating pairs/triples of features. As performed for $MAR4_{unifo}$, if the feature is continuous or ordinal, two groups are created using its median; if the feature is nominal, the existing categories are divided into two equally-sized groups. Then, for each feature, one of those groups is selected to have 2 × MR% of missing values.

## VI. CRITICAL ANALYSIS AND DISCUSSION

In this section, we provide a thorough analysis of some details that were left undiscussed in the original papers previously discussed, also referring to non-obvious issues that may arise in each implementation.

### A. MCAR UNIVA IMPLEMENTATIONS

Table 3 refers to some issues/restrictions in MCAR *univa* implementations. In what concerns $MCAR1_{univa}$, three main issues need to be considered:

- **Definition of MR:** By defining the MR over the entire dataset, the possible highest MR that is possible to simulate is dependent on the number of features comprised in the dataset. As an example, if dataset **X** has 2 features, the highest MR possible is limited to 50%, and ideally, should be lower, since, for 50%, $x_{miss}$ would be completely missing, given the $p\times$ MR adjustment.
- **Usage of Bernoulli trials:** To generate the missing values, $n$ Bernoulli trials are performed, each with probability of success $p$ = MR. According to the Law of Large Numbers (LLN), as the number of Bernoulli trials increases (as $n$, the number of patterns in **X** increases), the empirical probability of success (the real MR generated) will converge to the theoretical probability of success (the specified MR). As the name implies, the LLN applies when a large number of experiments is performed (large $n$). Therefore, for small datasets, there is no guarantee that the generated MR will coincide with the desired MR (it will be approximate, though not precise). As an example, for a desired MR of 30%, a certain run of MCAR generation could provide a real MR of 28% while another could return a real MR of 34%. Naturally, there is frequently a small bias in the generated missing percentages in several approaches, due to the rounding performed for the calculation of the number of missing positions to generate. However, this bias seems to be more significant when considering the usage of Bernoulli trials (for small datasets).
- **Correlation between features:** In all of the implementations by Twala [17] and Twala and Cartwright [24],

**TABLE 3.** Reviewed implementations for MCAR *univa* implementations: main characteristics and pitfalls.

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Twala et al. [17] | `MCAR1univa` | Random locations of $x_{miss}$ are derived from a Bernoulli distribution. $x_{miss}$ is the feature most correlated with the target class $t$. MR is defined for the whole dataset. MR for $x_{miss} = p \times n \times$ MR. | Bernoulli distribution may not guarantee the necessary missing rate. MR $< 100/p$. Correlation between features not addressed in the original paper. |
| Rieger et al. [18]<br>Xia et al. [19] | `MCAR2univa` | Random locations of $x_{miss}$ are deleted. $x_{miss}$ may be chosen randomly or by the researcher. MR definition may be chosen by the researcher. | MR $< 100/p$ if it is defined for the entire dataset and MR $< 100$ if its defined only for $x_{miss}$. |
| García-Laencina et al. [6] | `MCAR3univa` | Random locations of $x_{miss}$ are deleted. $x_{miss}$ can be chosen randomly or according to its relevance for classification (highest or lowest mutual information). MR is defined for a single feature. | MR $< 100$. Estimation of continuous probability density functions is challenging. |

**TABLE 4.** Reviewed implementations for MAR *univa* implementations: main characteristics and pitfalls.

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Twala et al. [17] | `MAR1univa` | Values of $x_{miss}$ corresponding to the lowest values in $x_{obs}$ are deleted. $x_{miss}$ is the feature most correlated with the target class $t$ and $x_{obs}$ is the feature most correlated with $x_{miss}$. MR is defined for the whole dataset. | MR $< 100/p$. Correlation between features not addressed in the original implementation. Computation of percentiles $k\%$ considered in the original implementation could be problematic for qualitative data. |
| Rieger et al. [18]<br>Xia et al. [19] | `MAR2univa` | Missingness on $x_{miss}$ depends on the ranks of $x_{obs}$. | MR $< 100/p$ if it is defined for the entire dataset and MR $< 100$ if its defined only for $x_{miss}$. MAR mechanism could be weakened in some situations. Random choice of $x_{obs}$ and $x_{miss}$ could weaken the consistency of experiments. |
| Rieger et al. [18] | `MAR3univa` | Values of $x_{miss}$ where corresponding values of $x_{obs}$ are equal to or higher than its median have a missing probability 9 times higher than the remaining values. | |
| | `MAR5univa` | For a total number of missing values $N$, $N/2$ locations of $x_{miss}$ are deleted for the highest values of $x_{obs}$ and $N/2$ for the lowest values. | |
| | `MAR4univa` | Values of $x_{miss}$ corresponding to the highest values of $x_{obs}$ are deleted. | MR $< 100/p$ if it is defined for the entire dataset and MR $<$ 100 if it is defined only for $x_{miss}$. Random choice of $x_{obs}$ and $x_{miss}$ could weaken the consistency of experiments. |

$x_{miss}$ is the feature most correlated with the class labels. Furthermore, in some approaches, there is also the need to define pairs of correlated features. In the original papers, Twala et al. consider datasets composed by both quantitative and qualitative features, yet the computation of the correlation between different types of features is not specified. Possible solutions to measure the correlation between different feature types are the computation of mutual information between features or the calculation of different coefficients according to each feature type (e.g. *Pearson* coefficient for two continuous features, *phi* coefficient for two binary features, *point-biserial* for a continuous and a binary feature, and so forth). The latter solution, however, would have to be looked at as an approximation, since there is no proper way to compare different coefficients.

$MCAR2_{univa}$ implementation allows the definition of MR for the entire dataset or for a single feature and depending on that choice, there are different restrictions to the allowed missing rates (Table 3). Regarding $x_{miss}$, it can be randomly chosen or defined by the researcher. To provide a consistent experimental setup, one could choose the same feature $x_{miss}$ to be missing at several MRs (e.g. 5, 10, 20%) and study the effects that higher MRs have in classification performance.

Choosing $x_{miss}$ according to the highest mutual information (MI) with the class labels $t$ ($MCAR3_{univa}$) might be problematic for quantitative/continuous features. The MI for two qualitative/categorical features is straightforward since the probability densities can be estimated using a histogram [22]. However, for quantitative/continuous features, the estimation of probability densities is more complicated. Frequent solutions include the discretization of continuous features [22] or applying Parzen-windows estimation [25], which is the method chosen for $MCAR3_{univa}$. The computation of Parzen windows can, however, be computationally expensive.

Among all approaches, $MCAR2_{univa}$ is an efficient method, straightforward to understand and implement, and thus we recommend it for standard MCAR *univa* experiments.

### B. MAR UNIVA IMPLEMENTATIONS

The limitations found for MAR *univa* implementations are summarized in Table 4. $MAR1_{univa}$ is based on finding a $k\%$ percentile of $x_{obs}$ to define a cut-off value: values of $x_{miss}$ lower than such cut-off are set missing.

Using this $k\%$ percentile might be problematic for nominal features (for which only mode applies) and ordinal features with several repeated values. Imagine $x_{obs}$ = `[1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3]`. If we were to consider $k = 50\%$, the percentile of $x_{obs}$ would be 3. However, setting values lower to 3 to missing would only return a $5/15 = 33\%$ missing rate.

In practice, the percentile should not be applied directly, and a simpler approach could be considered: deleting the

**TABLE 5.** Reviewed implementations for MNAR *univa* implementations: main characteristics and pitfalls.

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Twala et al. [17] | `MNAR1univa` | Lower values of $x_{miss}$ are deleted. $x_{miss}$ is the feature most correlated with the target class $t$. MR is defined for the whole dataset. | MR $< 100/p$. Correlation between features is not addressed in the original implementation. Computation of percentiles $k\%$ considered in the original implementation could be problematic for qualitative data. |
| Xia et al. [19] | `MNAR2univa` | Higher values of $x_{miss}$ are deleted. $x_{miss}$ can be chosen randomly or by the user. MR is defined for a single feature. | MR $<$ 100. Random choice of $x_{miss}$ could weaken the consistency of experiments. |

**TABLE 6.** Reviewed implementations for MCAR *unifo* implementations: main characteristics and pitfalls.

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Twala et al. [17] | `MCAR1unifo` | Random locations in each feature are derived from a Bernoulli distribution. All features will have missing data in the same percentage. | Bernoulli Distribution may not guarantee the necessary missing rate. MR $<$ 100. |
| Garciarena et al. [20]<br>Zhu et al. [21]<br>Pan et al. [22]<br>Ali et al. [23] | `MCAR2unifo` | Random locations $x_{i,j}$ are chosen to be missing. | Features may have very different percentages of missing data. High variability between runs of the algorithm. MR $<$ 100. |

lowest $k\%$ values, to guarantee that the desired missing rate is respected. For unordered features (nominal), however, the issue remains.

In $MAR2_{univa}$, higher ranks of $x_{obs}$ condition the missing positions in $x_{miss}$. According to (6), the missing positions should correspond to the highest ranks of $x_{obs}$. Nevertheless, (6) only defines the probability of each position in $x_{miss}$ to be deleted, which does not mean that a value with a low probability cannot be chosen to be deleted. From a pessimistic perspective, this means that values in $x_{miss}$ corresponding to both low and high ranks of $x_{obs}$ can be missing (although higher ranks are preferred) which would slightly break MAR assumption.

This issue is also shared by $MAR3_{univa}$, where $x_{obs}$ values higher than its median should define the missing positions in $x_{miss}$, although there is no guarantee that only $x_{miss}$ values corresponding to $x_{obs}$ values higher than the median are chosen. Besides, the objective of dividing two groups according to their median in $MAR3_{univa}$ is to create two approximately equally-sized groups, which might not be possible for ordinal features (similarly to $MAR1_{univa}$) and does not apply to nominal features. This could affect the $nG_1$ and $nG_2$ values and, in an extreme case, could lead to having the same probabilities for all values in $x_{miss}$, if $nG_1 = 9 \times nG_2$. An example would be a feature $x_{obs}$ = "Status" = `[1, 2, 2, 2, 2, 2, 2, 2, 2, 2]`, where all values would have the same probability (0.1) of generating missing positions in $x_{miss}$. This, however, traduces a MCAR mechanism, not MAR.

$MAR4_{univa}$ follows a standard approach for MAR generation, where the values of $x_{obs}$ are ordered and the $N$ highest values (according to the specified missing rate) are set missing. $MAR5_{univa}$, by generating $N/2$ missing values where $x_{obs}$ assumes its highest values and $N/2$ where it assumes the lowest, may create a rather blurred MAR mechanism for ordinal features. As an example, for $x_{obs}$ = `[1, 2, 2, 2, 2, 2, 3, 3, 3, 3]` a $MAR5_{univa}$ approach with MR = 60% would delete values of $x_{miss}$ corresponding to the subsets `[1, 2, 2]` (lowest) and `[3, 3, 3]` (highest). The MAR assumption would be hard to verify since it would seem that the values of $x_{obs}$ were not related to missing positions in $x_{miss}$. In turn, a 60% $MAR4_{univa}$ would delete values of $x_{miss}$ corresponding to the subset `[1, 2, 2, 2, 2, 2]` where the relation between $x_{obs}$ and $x_{miss}$ would be more clear: lower values in $x_{obs}$ condition the missingness of $x_{miss}$. Considering all approaches, $MAR4_{univa}$, although simple, seems the most robust. Nevertheless, for nominal features as the determining features ($x_{obs}$), both $MAR4_{univa}$ and $MAR5_{univa}$ would require some adjustments, since values cannot be ordered.

### C. MNAR UNIVA IMPLEMENTATIONS

Table 5 summarizes the characteristics of MNAR *univa* approaches. $MNAR1_{univa}$ suffers from the same restrictions as $MAR1_{univa}$, although the issues derived from the usage of the cut-off defined by the $k\%$ percentile may be attenuated by an ordering of values. After this modification, $MNAR1_{univa}$ and $MNAR2_{univa}$ are equivalent, except for three small differences: $MNAR1_{univa}$ chooses $x_{miss}$ as the most correlated with the class labels ($MNAR2_{univa}$ chooses randomly), $MNAR1_{univa}$ considers the lowest values of $x_{miss}$ while $MNAR2_{univa}$ chooses the highest and $MNAR1_{univa}$ considers MR for the entire dataset while $MNAR2_{univa}$ considers the MR for a single feature. $MNAR1_{univa}$ strives for consistency due to the choice of $x_{miss}$ while $MNAR2_{univa}$ strives for simplicity and flexibility: the definition of MR is not subjected to so much restrictions and the input of $x_{miss}$ can be easily adapted to consider a user-defined feature index. We therefore select $MNAR2_{univa}$ as the go-to implementation.

### D. MCAR UNIFO IMPLEMENTATIONS

The characteristics of MCAR *unifo* approaches are presented in Table 6. Given the use of Bernoulli trials, $MCAR1_{unifo}$ suffers from the same limitation of its *univa* analogous, where for small datasets (small $n$) the desired MR may not be

**TABLE 7. Reviewed implementations for MAR *unifo* implementations: main characteristics and pitfalls.**

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Garciarena et al. [20] | `MAR1unifo` | Values of the $n_{x_{miss}}$ features corresponding to the lowest values in $x_{obs}$ are set missing. $n_{x_{miss}}$ is specified by the researcher. | MR $\leq 100 \times n_{x_{miss}}/p$. Random choice of $x_{obs}$ and $x_{miss}$ may weaken the consistency of experiments. |
| Twala et al. [17] | `MAR2unifo` | Pairs of correlated features $\{x_{obs}, x_{miss}\}$ are defined. Values of $x_{miss}$ corresponding to the lowest values in $x_{obs}$ are deleted. For each pair, $x_{miss}$ is the feature most correlated with the target class $t$. | Correlation between features and formation of triples not addressed in the original paper. Computation of percentiles $k\%$ considered in the original implementation could be problematic for qualitative data. MR $< 50$. |
| Ali et al. [23] | `MAR3unifo` | Pairs of correlated features $\{x_{obs}, x_{miss}\}$ are randomly defined. Two groups in $x_{miss}$ are defined according to the median of $x_{obs}$. One of those groups is randomly chosen to have missing values. In each pair, $x_{miss}$ is randomly chosen. | Correlation between features and formation of triples is not addressed. Median may not always guarantee two equally-sized groups. MAR mechanism could be weakened in some situations. MR $< 25$. |
| Zhu et al. [21]<br>Pan et al. [22] | `MAR4unifo` | Random pairs of features $\{x_{obs}, x_{miss}\}$ are defined. For continuous or ordinal features, two groups in $x_{miss}$ are defined according to the median of $x_{obs}$; for nominal features, values are divided into two equally-sized groups and one is randomly chosen to have missing values. In each pair, $x_{miss}$ is randomly chosen. | MR $< 25$. In extreme scenarios, the median may not always guarantee two equally-sized groups for quantitative features or the necessary number values to delete. Division of qualitative values may also be problematic. MAR mechanism could be weakened in some situations. |

guaranteed. In $MCAR2_{unifo}$, since all $x_{i,j}$ are eligible to be missing, this approach generates a great amount of different missing datasets. Given the variability of possible missing datasets that can be generated (more than for $MCAR1_{unifo}$), it is fundamental that missing data experiments using it perform several runs.

These two approaches are rather different, therefore the choice of one will come down to the objectives and necessities of the experiments – $MCAR2_{unifo}$ is a popular approach [3], [7].

### *E. MAR UNIFO IMPLEMENTATIONS*

Table 7 summarizes the main characteristics and pitfalls of MAR *unifo* approaches. The flexibility given by $MAR1_{unifo}$ in what concerns the choice of the number of features to be missing leads to the restriction of possible missing rates according to (10).

$$MR \leq \frac{100 \times n_{x_{miss}}}{p} \text{ and } n_{x_{miss}} \leq p-1 \tag{10}$$

This means that, for a given number of missing features $n_{x_{miss}}$, it may not be possible to generate the desired MR and, conversely, that the number of chosen $n_{x_{miss}}$ may not be enough to guarantee the desired MR. As an example, consider a dataset $\mathbf{X}$ with $n = 303$ patterns and $p = 5$ features. To produce a MR of 60%, $n \times p \times \text{MR}/100 = 303 \times 5 \times 60/100 = 909$ values need to be missing. If only $n_{x_{miss}} = 2$ features are considered, that would mean that $909/2 = 455$ patterns would have to be missing in each feature, which is impossible. In this case, to guarantee that the MR would be respected, $n_{x_{miss}} \geq 4$ features should be considered.

$MAR2_{unifo}$ is subjected to the same issue as $MAR1_{univa}$ in what concerns the definition of $k\%$ percentiles. This issue may be surpassed in the same way as for $MAR1_{univa}$: instead of directly applying a cut-off value defined by $k$, one could consider the lowest $k\%$ values, to guarantee that the desired missing rate is achieved. A less obvious issue with $MAR2_{unifo}$ resides in the definition of MR and the creation of pairs/triples. Since the MR is defined for the entire dataset, the percentage of missing values in $x_{miss}$ needs to be adjusted accordingly: 2 × MR for pairs and 1.5 × MR for triples. Therefore, the maximum MR that can be specified to guarantee that the overall MR is achieved and that the $x_{miss}$ features are not completely deleted is $\text{MR} = 100/2 = 50\%$.

Regarding $MAR3_{unifo}$, using the median to define two groups and, more importantly, sampling missing values from only one of those groups may be problematic in some cases. Given the restriction of sampling from one of the groups, the MR generated in $x_{miss}$ is adjusted to 4 times higher (or 3 times higher for triples) so that the overall missing rate is respected. In some scenarios where $x_{obs}$ is qualitative, there might not be enough samples in one of the groups to choose from.

For instance, imagine a dataset composed of features $\{\text{``}Status\text{''}, \text{``}Age\text{''}\}$, where $x_{obs} = \text{``Status''}$, contains 1/2 values that encode "High" (70% of values) and "Low" (30% of values) status. Since the median of $x_{obs}$ will be 1, values lower or equal to 1 are put in one group (70%) and values higher than 1 are put in the other group (30%). If a MR of 20% is desired, then $4 \times \text{MR} = 40\%$ of missing values need to be generated in "Age" ($x_{miss}$). If the group "Status" = 2 is chosen to sample from, there are not sufficient samples to guarantee the desired MR. In another scenario, if "Status" values were coded as 1/0, then one of the groups would be empty since all values are lower or equal to the median: if that empty group was chosen to sample from, no missing data would be generated at all; if the other group (containing all data) is chosen instead, then 40% of the samples are randomly chosen considering all possible values. In this case, the MAR mechanism may not be respected given that missing values in "Age" would not be related to values of "Status": since all values are possible to choose from, this would more likely traduce a MCAR mechanism. Similarly to $MAR2_{unifo}$, some adjustments need to be performed for the MR in each $x_{miss}$ for pairs/triples. Accordingly, the maximum MR that can be specified is $\text{MR} = 100/4 = 25\%$.

**TABLE 8.** Reviewed implementations for MNAR *unifo* implementations: main characteristics and pitfalls.

| Publication | Algorithm | Description | Issues/Restrictions |
|---|---|---|---|
| Garciarena et al. [20] | MNAR1unifo | Lower values of $x_{miss}$ are deleted. $n_{x_{miss}}$ is defined by the researcher. | MR $\leq$ 100 $\times$ $n_{x_{miss}}/p$. Random choice of $x_{miss}$ may weaken consistency of experiments. |
| Twala et al. [17] | MNAR2unifo | Pairs/Triples of correlated features are defined. For each pair, the feature most correlated with the target class $t$ is chosen to be missing ($x_{miss}$): lower values of $x_{miss}$ are deleted. | Correlation between features and formation of triples is not addressed in the original paper. Computation of percentiles $k\%$ considered in the original implementation could be problematic for qualitative data. MR $<$ 50%. |
| Ali et al. [23] | MNAR3unifo | For each feature, two groups are defined according to its median. One of the groups is randomly chosen to have missing values. | MR $<$ 50%. Median may not always guarantee two equally-sized groups. MNAR mechanism could be weakened in some situations. |
| Zhu et al. [21]<br>Pan et al. [22] | MNAR4unifo | For continuous or ordinal features, two groups are defined according to its median; for nominal features, values are divided into two equally-sized groups. For each feature, one of these groups is randomly chosen to have missing values. | MR $<$ 50%. In extreme scenarios, the median may not always guarantee two equally-sized groups for quantitative features or the necessary number of values to delete. Division of qualitative values may also be problematic. MNAR mechanism could be weakened in some situations. |

$MAR4_{unifo}$ is the only approach that considers both quantitative and qualitative features. However, i) qualitative features with several repeated values can still weaken MAR assumption, as previously discussed and ii) the definition of two groups according to the median can still be problematic for quantitative features, if some values are repeated often. Besides, the generic strategy of creating two groups according to the median may not work well for high missing rates, since the adjustment of 4 × MR or 3 × MR that are required in each $x_{miss}$ may easily require the deletion of more values than the ones that exist in the defined groups.

Given the stronger restrictions in MR of $MAR2_{unifo}$, $MAR3_{unifo}$, and $MAR4_{unifo}$ implementations, we consider that $MAR1_{unifo}$ is perhaps the most adequate MAR *unifo* generation algorithm.

### F. MNAR UNIFO IMPLEMENTATIONS

MNAR *unifo* implementations are characterized in Table 8. Since they are very similar to their MAR *unifo* analogous, the same restrictions apply. $MNAR1_{unifo}$ suffers from the same restrictions as $MAR1_{unifo}$, due to the flexibility of choosing a given number $n_{x_{miss}}$ of missing features (11).

$$MR \leq \frac{100 \times n_{x_{miss}}}{p} \text{ and } n_{x_{miss}} \leq p \tag{11}$$

$MNAR2_{unifo}$ suffers from the same restrictions as $MAR2_{unifo}$, given that for MNAR, the pairs/triples of correlated features are also defined and, therefore, the respective adjustments to the MR need to be applied.

$MNAR3_{unifo}$ and $MNAR4_{unifo}$ do not require the formation of pairs/triples since all the features will have missing values. Nevertheless, due to the formation of two groups for each feature, the MR needs to be adjusted as well. For a specified MR, each feature $x_{miss}$ needs to have MR% of missing values. However, since two groups are defined for each feature (with approximately 50% of data, which is the objective of using the median) and only one of those groups is used to generate missing values, then the maximum possible MR is 50%. As in previous approaches, the use of the median might be problematic in some scenarios. First, it may not guarantee two equally-sized groups and, therefore, the desired MR might not be achieved; secondly, and especially in the case of $MNAR3_{unifo}$, for qualitative features with several repeated values, the MNAR assumption may be weakened, as explained for MAR mechanism.

Again, given the stronger restrictions in MR of $MNAR2_{unifo}$, $MNAR3_{unifo}$ and $MNAR4_{unifo}$ implementations, we consider that $MNAR1_{unifo}$ is perhaps the most adequate MNAR *unifo* generation algorithm.

## VII. DOMAIN-BASED MISSING DATA GENERATION APPROACHES

The implementations presented in the previous sections are rather generic approaches to missing data generation. They were developed for general domains, with no particular focus on the peculiarities of a given domain and without assuming any *apriori* knowledge of the domain (e.g. known relationships between features in the study). However, some missing data generation approaches found in the literature are adapted to the domain in question. In this section, we review some domain-specific approaches to missing data generation. Some, although uncommon, may be generalized to different domains; others are not generalizable but may contain interesting details to consider for some real-world domains (e.g. healthcare domains).

Song and Shepperd [26] focus on evaluating imputation methods for small project effort data sets. In this domain, MAR data is generated according to the size of the project. First, records are ordered by project size; then, the dataset is divided into 4 parts with different percentages of missing data: for each part $d$, its missing percentage is proportional to $\frac{M_d}{\sum_{d=1}^{4} M_d}$ × MR, where $M_d$ is the mean of project size of the $d^{\text{th}}$ part.

Josse *et al.* [27] use synthetic data to generate two different MAR scenarios, "MAR easy" and "MAR difficult" for a simulated dataset comprising 9 features that could be divided into two blocks of correlated features: $\{x_1, x_2, x_3, x_4, x_5\}$ and $\{x_6, x_7, x_8, x_9\}$. Then, "MAR easy" would consist of deleting values of $x_2$ to $x_5$ according to values of $x_1$ and deleting

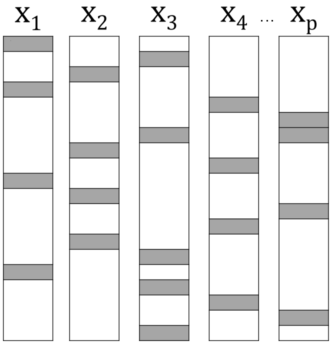


**FIGURE 11.** Missing data pattern of the MCAR implementation by Nanni *et al.* [30].

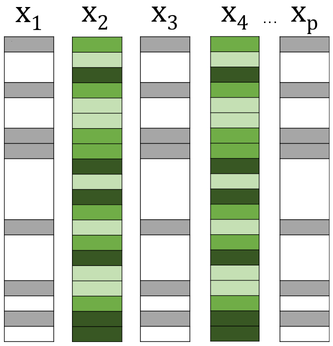


**FIGURE 12.** Missing data pattern of the MuOv implementation by Garciarena and Santana [20].

values of $x_6$ to $x_8$ according to values of $x_9$. This traduces a situation where the missing values are easier to recover given the known existing correlation between features. "MAR difficult" worked by deleting values of $x_6$ to $x_9$ according to values of $x_1$ and deleting values of $x_1$ to $x_5$ according to values of $x_9$, so that the available information to predict missing values is very limited.

Johansson and Karlsson [28] focus on strategies to handle missing values in clinical data. A pharmacokinetic model was used to generate a synthetic dataset where missing values were generated in feature "Sex". For MCAR, values of "Sex" were randomly deleted; for MAR missing values in "Sex" were generated according to the "Weight" of the subjects and finally, for MNAR, missing values in "Sex" were deleted for male subjects.

Olsen *et al.* [29] study the effects of handling missing data in clinical trials of knee osteoarthritis. Missing data was generated in two MNAR scenarios: Scenario A, where the probability of missing data was dependent on changes of pain, physical function and patient's global assessment, and Scenario B, where the missingness was dependent on type of treatment and consequent effects.

Nanni *et al.* [30] focus on discovering an imputation method that would perform well in medical domains. Thus, authors generate MCAR data in a different fashion: instead of generating MR% of missing values in each feature or in the whole dataset directly (deleting MR% of $x_{i,j}$ elements), the missing values are generated in each pattern $\mathbf{x}_i$. In other words, each pattern $\mathbf{x}_i$ will have MR% of missing values, where different features can be missing for different patterns (Figure 11). This translates a context where all patients have at least one missing observation.

Deb and Liew [31] study missing value imputation for the analysis of traffic accident data and generate missing values in a similar way to Nanni *et al.* [30]: missing values are generated by pattern, rather than by feature. This generation method follows from the research of Rahman and Islam [32], [33] and considers four main configuration types: *Simple*, *Medium*, *Complex* and *Blended*. In the *Simple* generation, each pattern $\mathbf{x}_i$ has at most one missing value; in *Medium* generation, each pattern $\mathbf{x}_i$ has a minimum of 2 missing values and at most 50% missing values and in *Complex* generation, each pattern $\mathbf{x}_i$ has between 51% and 80% missing values. Finally, *Blended* generation considers a mixture of the remaining types – 25%, 50% and 25% of patterns according to *Simple*, *Medium* and *Complex* generation types, respectively.

Furthermore, two different models for missing data generation are used: Uniformly Distributed (UD) and Overall models. In the UD model, it is guaranteed that all features have the same amount of missing values, whereas in the Overall model missing values can be dispersed across several features (in the worst-case scenario, they can all appear in a single feature).

Garciarena and Santana [20], as mentioned in Section V-C, propose another version to generate MNAR data, called MuOv (Missing depending on unobserved Variables). MuOv represents a MNAR scenario where the probability of missing values in a feature is related to some other feature that was not considered in the study. In this case, $N$ patterns are randomly chosen to be missing (according to the desired MR) and their values on each feature to be missing are deleted (Figure 12). Although MuOv does not consider the application to a specific domain, we have included it here since it is rather an uncommon MNAR approach, as previously discussed.

Valdiviezo and Van Aelst [34] introduced missing values in real-world datasets according to different mechanisms and schemes. In general, two schemes are followed for each mechanism: either considering all features (*first scheme*) or considering only one-third of features, which are randomly chosen (*second scheme*). Regarding MCAR mechanism, the *first scheme* inserts MR% of missing values in each feature while in the *second scheme*, since only one-third of features will have missing data, each of those features will have 3 × MR% of missing values. In the original paper, this adjustment is not mentioned, but we have decided to discuss it so that the overall MR is respected.

In MAR mechanism, the *first scheme* randomly selects one feature to be the determining feature, $x_{obs}$, and the remaining $p-1$ features will have their values missing according to the values of $x_{obs}$. To that end, the values of $x_{obs}$ are transformed into probabilities by a logistic function, and the missing locations for the remaining features are sampled according to such probability.

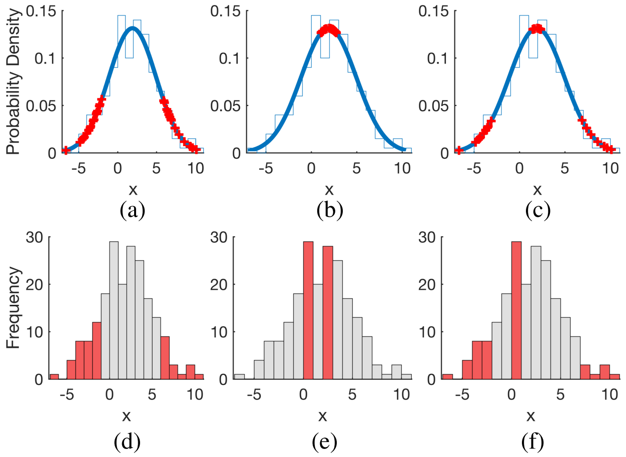


**FIGURE 13.** Strategies for missing data generation: $T_1$ to $T_3$ are *pdf*-based methods while $T_4$ to $T_6$ are *freq*-based methods. (a) $T_1$: *pdf*-outer. (b) $T_2$: *pdf*-inner. (c) $T_3$: *pdf*-both. (d) $T_4$: *freq*-outer. (e) $T_5$: *freq*-inner. (f) $T_6$: *freq*-both.

Finally, in MNAR mechanism, the *first scheme* deletes the highest or lowest values of each feature in the dataset, while the *second scheme* proceeds in the same way but only for one-third of features.

Soares *et al.* [35] study how different methods behave when imputing data from different continuous distributions. To that end, each feature is fitted against a comprehensive set of continuous distributions and missing values are generated according to 7 distinct methods, $T_1$ to $T_7$. Method $T_7$ is a standard MCAR approach ($MCAR2_{univa}$), where the same amount of missing values are randomly inserted in each feature. The remaining methods are MNAR approaches, where the missing values are removed according to each feature's probability density functions or frequency histograms. Methods $T_1$ to $T_3$ are *pdf-based* while methods $T_4$ to $T_6$ are *freq-based*. For each method, three different scenarios are considered: removing from the outer areas, inner areas or both. Outer and inner areas correspond to low and high values of the *pdf* and frequency histogram, respectively (Figure 13).

## VIII. DISCUSSION

Overall, we may divide the issues of reviewed approaches into three different types: Theoretical flaws, Empirical flaws and Experimental Setup hazards. *Theoretical flaws* refer to design flaws in the approaches: problems that may arise in some of the key ideas of the approach. *Empirical flaws* refer to some issues that may occur not (solely) due to the rationale behind each approach, but generated by specific conditions that may arise in some domains (e.g. different feature types), often discussed throughout the paper. Finally, *Experimental Setup hazards* are not considered *flaws* inherent to the approaches *per se*, but refer to some details that should be taken into account: they are considered *hazards* in the sense that they are risks, but can easily be surpassed by a careful experimental design.

- *Theoretical flaws:*
  - ○ **Usage of Bernoulli trials:** For datasets with a small number of patterns (small $n$), Bernoulli trials may not provide the desired MR. To surpass this issue several algorithms use random permutations of $x_{i,j}$ positions instead.
  - ○ **Definition of pairs/triples and consequent MR adjustments:** On one hand, defining pairs/triples of features is an interesting approach since we guarantee that there is a relation between the features. In MAR and MNAR, for each missing feature, there is another highly correlated with it (completely observed) that, in theory, possesses information that may be relevant when imputing the missing values. However, defining these pairs/triples may condition the MR greatly, due to the necessary adjustments: depending on the implementation, the MR may be limited from less than 50% to less that 25%.
- *Empirical flaws:*
  - ○ **Usage of the median to define groups:** Using the median generally aggravates the MR restrictions, especially for MAR *unifo* implementations. Furthermore, if the dataset comprises qualitative features, the use of the median can, in some situations, weaken the mechanisms or fail to provide the specified MR (e.g. $MAR3_{univa}$, $MAR3_{unifo}$, $MAR4_{unifo}$, $MNAR3_{unifo}$, $MNAR4_{unifo}$, among others).
  - ○ **Usage of cut-off values defined using percentiles:** Defining a cut-off value and deleting values accordingly might fail to provide the desired MR, especially if qualitative features are at state, as explained throughout the paper. Among all implementations, cut-off values based on percentiles are only considered in Twala [17], which could be replaced by a sorting of the values to guarantee that the necessary MR is respected. Nevertheless, the sorting requires that a feature can be ordered, which is not always the case.
- *Experimental Setup hazards:*
  - ○ **Random choice of determining and missing features:** If we consider a typical experimental setup where $n$ datasets are chosen to generate missing values, one important aspect is to make the experiments as consistent as possible. As an example, consider a dataset $\mathbf{X}$ where MAR values are generated in MRs of 10%, 20%, 30% and so on. If missing values are generated according to $MAR1_{unifo}$, for instance, where the determining and missing features are randomly chosen, there are several factors (besides the increase of the MR) that affect the final results. These type of assumptions and limitations need to be established *apriori*, according to the objectives of the experiments. In some cases, the presented domain-based approaches might be worthy of consideration (see [30]), adapting the missing value generation to the context and objectives of the study.
  - ○ **Variability of generated missing datasets:** In some cases, especially in MCAR approaches,

the possibilities of obtaining different outcomes is enormous and, therefore, several runs should be performed. As an example, two different runs of $MAR2_{unifo}$ might provide datasets with different "difficulty degrees" for imputation algorithms. Nevertheless, this is not an issue of the approach *per se* but should be bypassed by the design of experiments.

## IX. CONCLUSIONS AND POTENTIAL RESEARCH DIRECTIONS

This manuscript reviews a considerable number of missing data generation approaches, for different configurations (univariate and multivariate) and missing data mechanisms (MCAR, MAR, and MNAR). Their limitations are discussed from a theoretical and empirical view and some modifications are suggested in order to surpass them. Additionally, we refer some less common approaches – herein named "domain-based" approaches – in order to illustrate existing missing data generation approaches in specific contexts.

The *theoretical flaws* may compromise/constraint the possible MRs to generate; nevertheless, this problem is easy to diagnose and, although the desired percentage of missing values may not be achieved, there is no risk of breaking the assumptions regarding the missing mechanisms. Regarding the identified *empirical flaws*, it is important to state that they are mostly related to the existence of qualitative features with no order (nominal features), which is very common across several domains [3], [36]. This is the most challenging topic to solve in related work and is most often neglected. With the exception of Zhu *et al.* [21] and Pan *et al.* [22], which distinguish between ordered and nominal features, no other work refers to this issue. This limitation becomes more evident when using the median or percentiles/quantiles, which require that the features have an order, although in any implementation that requires values to be ordered (independently of the use of median or percentiles), this problem exists. These *empirical flaws* are more serious since they may bias the missing mechanism. The *experimental setup hazards* are unrelated to the described approaches, but they might be induced inadvertently by the researcher during the study. Therefore, they will not affect certain aspects of the implementation (faulty MR rate, broken missing mechanism), but they may compromise the derived insights for certain implementations, if there is not a careful experimental design (e.g overlooking the stochastic process inherent to the MCAR *unifo* approaches).

Domain-based approaches are mainly developed in order to adapt to given contexts: they arise when there is a need to study specific situations/properties in data (Josse *et al.* [27] and Soares *et al.* [35]) or to map known relationships in data (Johansson and Karlsson [28] and Olsen *et al.* [29]), or they may also reflect the reality of certain domains, such as healthcare domains (Nanni *et al.* [30]), software management (Song and Shepperd [26]) and traffic data (Deb and Liew [31]). A standard approach in this case is to generate missing values per pattern, rather than per feature. This is a way of traducing the reality in these domains: as an example, in medical datasets, it is not expected that certain features are absent for all patients, but instead, that several patients have absent observations in some features [3]. Although some of these approaches are not generalizable, we have decided to present them since they represent valid approaches in certain contexts and might inspire other approaches for similar domains.

Finally, we shall refer to some potential research directions in the field:

- **Generating MAR and MNAR with nominal features:** The definition of appropriate strategies to generate MAR and MNAR data with nominal features would be important, since most strategies proposed so far may fail under certain circumstances;
- **Generating MAR through data modeling:** In the reviewed works, MAR either makes use of one *determining* feature, or pairs of features where one in the pair is the *determining* feature. Future research could explore the effects of generating MAR via the combination of all features in data (except for the missing features);
- **Investing in software development:** Nowadays, a great number of statistical software (SPSS, R, MatLab) considers the development of models with missing data and procedures for MD imputation. Nevertheless, strategies for MD generation are most often neglected;
- **Experimenting over real-world datasets:** Investigating the reliability and consistency of the methods outlined in this work on a large benchmark of real-work datasets (available from UCI or Kaggle repositories [37], [38]), comprising different domains, number of samples, number (and type) of features and distributions could prove beneficial to the literature.

## REFERENCES

[1] P. H. Abreu *et al.*, "Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data," in *Proc. 13th Medit. Conf. Med. Biol. Eng. Comput.*, 2014, pp. 1366–1369.

[2] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Comput. Biol. Med.*, vol. 59, pp. 125–133, Apr. 2015.

[3] M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *J. Biomed. Inform.*, vol. 58, pp. 49–59, Dec. 2015.

[4] T. D. Little, T. D. Jorgensen, K. M. Lang, and E. W. G. Moore, "On the joys of missing data," *J. Pediatric Psychol.*, vol. 39, no. 2, pp. 151–162, 2013.

[5] D. C. Howell, "The treatment of missing data," in *The Sage Handbook of Social Science Methodology*. London, U.K.: Sage, 2007, pp. 208–224.

[6] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.

[7] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, "Influence of data distribution in missing data imputation," in *Proc. Conf. Artif. Intell. Med. Eur.* Vienna, Austria: Springer, 2017, pp. 285–294.

[8] J. W. Graham, P. E. Cumsille, and E. Elek-Fisk, "Methods for handling missing data," in *Handbook of Psychology*. Hoboken, NJ, USA: Wiley, 2003, pp. 87–114.

[9] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annu. Rev. Psychol.*, vol. 60, pp. 549–576, Jan. 2009.

[10] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Classifying patterns with missing values using multi-task learning perceptrons," *Expert Syst. Appl.*, vol. 40, no. 4, pp. 1333–1341, 2013.

[11] S. Nakagawa, "Missing data: Mechanisms, methods and messages," in *Ecological Statistics: Contemporary Theory and Application*. Oxford, U.K.: Oxford Univ. Press, 2015, pp. 81–105.

[12] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[13] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 333. Hoboken, NJ, USA: Wiley, 2014.

[14] C. K. Enders, *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press, 2010.

[15] S. van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL, USA: CRC Press, 2012.

[16] P. D. Allison, *Missing Data*, vol. 136. Newbury Park, CA, USA: Sage, 2001.

[17] B. Twala, "An empirical comparison of techniques for handling incomplete data using decision trees," *Appl. Artif. Intell.*, vol. 23, no. 5, pp. 373–405, 2009.

[18] A. Rieger, T. Hothorn, and C. Strobl, "Random forests with missing values in the covariates," Dept. Statist., Univ. Munich, Munich, Germany, Tech. Rep. 79, 2010.

[19] J. Xia *et al.*, "Adjusted weight voting algorithm for random forests in handling missing values," *Pattern Recognit.*, vol. 69, pp. 52–60, Sep. 2017.

[20] U. Garciarena and R. Santana, "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers," *Expert Syst. Appl.*, vol. 89, pp. 52–65, Dec. 2017.

[21] B. Zhu, C. He, and P. Liatsis, "A robust missing value imputation method for noisy data," *Appl. Intell.*, vol. 36, no. 1, pp. 61–74, 2012.

[22] R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information," *Appl. Intell.*, vol. 43, no. 3, pp. 614–632, 2015.

[23] N. A. Ali and Z. M. Omer, "Improving accuracy of missing data imputation in data mining," *Kurdistan J. Appl. Res.*, vol. 2, no. 3, pp. 66–73, 2017.

[24] B. Twala and M. Cartwright, "Ensemble missing data techniques for software effort prediction," *Intell. Data Anal.*, vol. 14, no. 3, pp. 299–331, 2010.

[25] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[26] Q. Song and M. Shepperd, "A new imputation method for small software project data sets," *J. Syst. Softw.*, vol. 80, no. 1, pp. 51–62, 2007.

[27] J. Josse, M. E. Timmerman, and H. A. L. Kiers, "Missing values in multilevel simultaneous component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 129, pp. 21–32, Nov. 2013.

[28] Å. M. Johansson and M. O. Karlsson, "Comparison of methods for handling missing covariate data," *AAPS J.*, vol. 15, no. 4, pp. 1232–1241, 2013.

[29] I. C. Olsen, T. K. Kvien, and T. Uhlig, "Consequences of handling missing data for treatment response in osteoarthritis: A simulation study," *Osteoarthritis Cartilage*, vol. 20, no. 8, pp. 822–828, 2012.

[30] L. Nanni, A. Lumini, and S. Brahnam, "A classifier ensemble approach for the missing feature problem," *Artif. Intell. Med.*, vol. 55, no. 1, pp. 37–50, 2012.

[31] R. Deb and A. W.-C. Liew, "Missing value imputation for the analysis of incomplete traffic accident data," *Inf. Sci.*, vol. 339, pp. 274–289, 2016.

[32] M. G. Rahman and M. Z. Islam, "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques," *Knowl.-Based Syst.*, vol. 53, pp. 51–65, Nov. 2013.

[33] M. G. Rahman and M. Z. Islam, "FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis," *Knowl.-Based Syst.*, vol. 56, pp. 311–327, Jan. 2014.

[34] H. C. Valdiviezo and S. Van Aelst, "Tree-based prediction on incomplete data using imputation or surrogate decisions," *Inf. Sci.*, vol. 311, pp. 163–181, Aug. 2015.

[35] J. P. Soares, M. S. Santos, P. H. Abreu, H. Araújo, and J. Santos, "Exploring the effects of data distribution in missing data imputation," in *Proc. Int. Symp. Intell. Data Anal.* Hertogenbosh, The Netherlands: Springer, 2018, pp. 251–263.

[36] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting breast cancer recurrence using machine learning techniques: A systematic review," *ACM Comput. Surv.*, vol. 49, no. 3, 2016, Art. no. 52.

[37] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[38] *Kaggle Datasets*, Kaggle Inc, San Francisco, CA, USA, 2018.

**MIRIAM SEOANE SANTOS** received the master's degree in biomedical engineering from the University of Coimbra, in 2014, where she is currently pursuing the Ph.D. degree in information science and technology. She is also a member of the Centre of Informatics and Systems, University of Coimbra. Her research interests include pattern recognition problems, imbalanced and missing data, and personalized medicine in oncology.

**RICARDO CARDOSO PEREIRA** received the bachelor's degree from the Coimbra Institute of Engineering and the master's degree in informatics engineering from the University of Coimbra, where he is currently pursuing the Ph.D. degree. He is also an Invited Assistant Teacher with the University of Coimbra. He also teaches at the Coimbra Institute of Engineering. His main research interests include computational intelligence and data processing, with a special emphasis on the missing data field.

**ADRIANA FONSECA COSTA** received the master's degree in biomedical engineering from the University of Coimbra, in 2018. She specialized in clinical informatics and bioinformatics. Her current research interests include data preprocessing problems, such as missing data, and deep learning applications to data imputation.

**JASTIN POMPEU SOARES** was born in Mönchengladbach, Germany. He received the master's degree in electrical and computer engineering from the University of Coimbra, in 2017. His research interests include preprocessing problems in knowledge discovery in databases and preprocessing and classification problems in clinical applications.

**JOÃO SANTOS** received the M.Sc. degree in physics (optoelectronics and lasers) and the Ph.D. degree in physics (condensed matter). He concluded the Residency in medical physics, in 2005, with the Portuguese Institute of Oncology of Porto (IPOPFG, EPE). He is currently a Medical Physicist Expert and a Coordinator of the Medical Physics, Radiobiology and Radiation Protection Group, IPOPFG Research Centre, and an Affiliate Professor with the University of Porto.

**PEDRO HENRIQUES ABREU** received the degree in informatics engineering and the Ph.D. degree in soccer teams modeling from the University of Porto, in 2006 and 2011, respectively. He is currently an Assistant Professor with the Department of Informatics Engineering, University of Coimbra. He has authored more than 60 publications in international conferences and journals. His research interests include medical informatics and personal healthcare systems applied to oncology.