Missing Data Imputation for Categorical Variables

Jaroslav Horníček[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Hana Řezanková[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 20.1.2022, **Accepted** (reviewed) 9.2.2022, **Published** 16.9.2022

Abstract

Dealing with missing data is a crucial part of everyday data analysis. The IMIC algorithm is a missing data imputation method that can handle mixed numerical and categorical datasets. However, the categorical data are crucial for this work. This paper proposes the new improvement of the IMIC algorithm. The two proposed modifications consider the number of categories in each categorical variable. Based on this information, the factor, which modifies the original measure, is computed. The factor equation is inspired by the Eskin similarity measure that is known in the hierarchical clustering of categorical data. The results show that as the missing value ratio in the dataset grows, better results are achieved using the second modification. The paper also shortly analyzes the advantages and disadvantages of using the IMIC algorithm.

Keywords	***DOI***	***JEL code***
IMIC algorithm, missing value imputation, categorical variables	*https://doi.org/10.54694/stat.2022.3*	*C38, C40, C80*

INTRODUCTION

The missing value imputation problem can be frequently encountered in natural and social sciences and technology. NASA uses missing value imputation when reconstructing images sent from outer space because it is not technologically possible to transfer every image pixel without information loss. On the other hand, social scientists may use this method in a survey to compensate for the reluctance of the respondents to answer questions. The correct imputation of missing values in such cases is crucial and affects the quality of the final research. Statistical analysis methods often require the information inherent in data to be complete (no missing values). Otherwise, the methods fail.

Before the basic methods for working with missing values are introduced, the basic terminology will be mentioned. There are three natural mechanisms that can cause incomplete data to occur. Namely, it is MCAR, MAR, and MNAR, described by Rubin (1976), Rubin and Little (2002), or Baraldi and Enders (2010). MCAR (missing completely at random) occurs when the data are missing randomly without any observable pattern. MAR (missing at random) happens when the missing values of one variable are dependent on another variable, e.g. with decreasing attained level of respondent education, there are more

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. Corresponding author: e-mail: horj31@vse.cz.
[2] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: hana.rezankova@vse.cz.

missing values in the variable for respondent income. Finally, MNAR (missing not at random) occurs when the observed values depend on each other, e.g., the respondents who have problems with alcohol might be less willing to answer in a survey on alcoholism (Petrúšek, 2015).

The simplest solution to dealing with missing values is simply removing the whole incomplete multidimensional observations. However, this reduces the number of observations and affects the randomness of the sample selection, which is mentioned by Azar (2002) or de Leeuw et al. (2003). In the case of pairwise statistical methods (e.g. correlation analysis), we do not have to remove every incomplete multidimensional observation. Pairwise methods can, in some cases, avoid the problem of deleting the incomplete observations from the sample, but it does not solve it.

One of the simplest methods of imputing the missing values is the replacement of the missing values with their mean. This approach may not change the mean value of the variable but can significantly affect the variability of the result. This method can safely be used only in cases where the mechanism of missing values is MCAR, as Baraldi and Enders (2010) explain.

A more advanced method of missing value imputation uses a regression function. However, even this method can significantly affect the variability of the sample. In practice, stochastic regression is often used. A random error term is added to the individual predicted values in such a case. This ensures that the imputed values do not strictly copy the given regression function and artificially create variability of the result, which is desirable in most cases, as Baraldi and Enders (2010) point out.

Apart from regression, there are many simple methods for imputing the missing values. These methods are based on simple linear models and other prediction algorithms of machine learning. Significant improvement was only achieved by introducing the multiple imputation method (Rubin, 1987) and the method based on the maximal likelihood estimation (Allison, 2012). However, even with these methods, we cannot avoid some inaccuracy in the estimation of the true values in case the missing data were created by the MNAR mechanism. Nevertheless, the estimation is demonstratively better than in the case of the simple methods, as Schafer and Graham (2002) say.

In their simplest form, the multiple imputation methods randomly select a subset of the original dataset and conduct a regression analysis on it. From each (stochastic) regression function obtained this way, missing values can be predicted. The final value is then a result of calculating the mean of these values. The procedure can be modified by selecting several subsequent stochastic regressions, where the correlation estimation and the mean from the previous step are used to calculate the new regression coefficients of each new regression, as Baraldi and Enders (2010) explain.

Methods based on the maximal likelihood are built on a complex mathematical background, which is out of the scope of this article. Both advanced methods (multiple imputation and maximal likelihood estimation) are currently recommended for handling in missing values. Unfortunately, they can mostly work for quantitative data only.

Regardless of the multiple imputation method for categorical data introduced by Akande et al. (2017), there is generally no missing value imputation method for categorical data, which would be demonstratively better than any other. However, several papers (Sulis and Porcu, 2008; Ferrari et al., 2011; Wu et al., 2012; Pecáková, 2014; Stavseth et al., 2019) study this topic. Worthy of mention is the method based on the rough sets theory. It is a hierarchical method, which looks for the nearest pairs within a set of observations with the help of a specially defined metric. If a pair of observations contains one missing and one non-missing value in a certain variable, the missing value is replaced by a non-missing value in this pair. An important concept based on the rough sets theory is a so-called extended tolerance relation, explained by Nguyen et al. (2013), which can measure the similarity between a complete and an incomplete observation. As a sufficient explanation, we can say that any two identical sets would be in the equivalence relation. If a value from a set is deleted and called missing, the equivalence relation would no longer exist, but a tolerance relation still exists.

Feng et al. (2011) introduced the IMIC algorithm. The IMIC is a missing data imputation method that can handle both categorical and numerical variables. This algorithm does not need a set of predictor variables without missing values for the prediction of missing values in another variable. Due to the hierarchical clustering, every single variable in the dataset can contain missing values, and the IMIC algorithm fills in all unknown values in one run of the iteration process.

The IMIC algorithm can easily handle missing values in multiple variables in an incomplete dataset. It can be easily used by an inexperienced user. These advantages make the algorithm very promising. On the other side, it is very time-consuming because the algorithm computes similarity measures between each pair of observations in hierarchical clustering. Hence efficient implementation is crucial.

This paper proposes the new improvement of the IMIC algorithm. The two proposed modifications consider the number of categories in each categorical variable. Based on this information, the factor, which modifies the original measure, is computed. The factor equation is inspired by the Eskin similarity measure (Eskin et al., 2002) that is known in the hierarchical clustering of categorical data (Šulc and Řezanková, 2014; Cibulková et al., 2021). The results show that as the missing value ratio in the dataset grows, better results are achieved using the modification.

1 THE MAIN PRINCIPLE OF THE IMIC ALGORITHM

The IMIC algorithm utilizes hierarchical clustering. At the beginning of the process, each observation X_i (the vector of the variable values) is an isolated cluster; $X_i \subset X$, $i \in \{1, 2, ..., n\}$, where n is the number of observations in the dataset X. The cluster is a name either for two or more observations joined together or for one single observation (one element cluster).

In the case of categorical variables only, the algorithm computes $ISMD_C$ (Incomplete Set Mixed Dissimilarity in Categorical attributes) between two clusters in the rth step of the algorithm as

$$ISMD_C(X_i^r, Y_j^r) = \frac{\left|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) = \emptyset\}\right|}{\sqrt{\left|X_i^r\right| + \left|Y_j^r\right|} \cdot \left|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}\right|}, \tag{1}$$

where X_i^r, Y_j^r are two different clusters $X_i^r \subset X$ and $Y_j^r \subset X$, $k \in \{1, 2, ..., q\}$, q is the number of categorical variables, the operator $|\cdot| : A \to N$ returns the number of elements in the set A and $|X_i^r|$ returns the number of observations in the cluster X_i^r, symbol $\emptyset$ represents empty set, and $s_k(X_i^r, Y_j^r)$ is defined as

$$s_k(X_i^r, Y_j^r) = \begin{cases} s_k(Y_j^r), & s_k(X_i^r) = * \wedge s_k(Y_j^r) \neq * \\ s_k(X_i^r), & s_k(X_i^r) \neq * \wedge s_k(Y_j^r) = * \\ s_k(X_i^r) \cap s_k(Y_j^r), & s_k(X_i^r) \neq \emptyset \wedge s_k(Y_j^r) \neq \emptyset \end{cases}, \tag{2}$$

where the symbol * represents a missing value, symbol $\cap$ represents the set intersection operator, $\wedge$ represents logical "and", and $s_k(X_j^r)$ is defined as

$$s_k(X_i^r, Y_j^r) = \begin{cases} v_{a_{kp}}, & (\exists x_i \in X_j^r(a_k(x_i) = v_{a_{kp}}) \wedge (\forall x_i \in X_j^r)(a_k(x_i) = v_{a_{kp}} \vee a_k(x_i) = *) \\ *, & (\forall x_i \in X_j^r)(a_k(x_i) = *) \\ \emptyset, & (\forall p)(\exists x_i \in X_j^r)(a_k(x_i) \neq v_{a_{kp}} \vee a_k(x_i) \neq *) \end{cases}, \tag{3}$$

where $\vee$ represents logical "or", and $v_{a_{kp}}$ is the value of the kth variable and the pth value of all c_k unique values of the kth variable ($p \in \{1, 2, ..., c_k\}$), $a_k(x_i)$ represents the value of the kth variable and the ith cluster. Based on the algorithm of Feng et al. (2011), the set of s_k for one cluster is denoted as CS feature.

In the first step ($r = 0$), every value of $s_k(X_j^r)$ is set equal $v_{a_{kp}}$ or * based on the true value in the kth variable and the ith cluster trivially (the cluster is one single observation in the case of $r = 0$). After that, the $ISMD_C(X_i^r, Y_j^r)$ is computed between every two clusters. In the case of categorical variables only, every pair X_i^r, Y_j^r with minimal $ISMD_C(X_i, Y_i)$ are added together, so these two clusters X_i^r, Y_j^r rare joined to the new cluster $X_{i,j}{}^r = X_f^{(r+1)}$, where $f \in \{1, 2, ..., q - t\}$, where t means number of cluster pairs added together in the rth step.

When the one step of clustering based $ISMD_C$ is finished, the algorithm tries to replace missing values in the cluster with one or more missing value * as

$$a_k(X_i^r) = \begin{cases} v_k, & v_k \in s_k(X_i^r) \wedge v_k \neq * \\ *, & v_k \in s_k(X_i^r) \wedge v_k = * \\ Mode_k(X_i^r), & s_k(X_i^r) = \emptyset \end{cases}, \tag{4}$$

After that, we set $r = r + 1$ and proceed new iteration until no missing values are present or $r = n$.

The $ISMD_C$ is increasing as the clusters are growing. The $ISMD_C$ in Formula (1) can be understood as a ratio of different values to same values in each variable of the new potential cluster. In other words, two clusters will be joined more likely if the values in compared variables are the same.

For a better understanding of the algorithm, there is a small example. Assume that the small set of binary data is given

$$U_0 = (\{X_1^0\}, \{X_2^0\}, \{X_3^0\}, \{X_4^0\}) = (\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}) = \begin{pmatrix} \{0 & 1 & 1 & 1 & 0\} \\ \{0 & * & 1 & 1 & 0\} \\ \{1 & 0 & 0 & * & 1\} \\ \{1 & 0 & 1 & 0 & 1\} \end{pmatrix},$$

where U_0 represents the initial set of multidimensional observations of the four observations (clusters) ($\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$). As can be seen, each of these clusters consists of the values of five variables. It is evident that the second observation in the second variable and the third observation in the fourth variable contain a missing value.

Firstly, the algorithm computes the set of s_k based on Formula (3). For the first cluster $X_1{}^0$, the CS feature will be equal $CS(x_1) = \{\{0\},\{1\},\{1\},\{1\},\{0\}\}$, for the second cluster $CS(x_2) = \{\{0\},\{*\},\{1\},\{1\},\{0\}\}$, for the third cluster $CS(x_3) = \{\{1\},\{0\},\{0\},\{*\},\{1\}\}$, and for the fourth cluster $CS(x_4) = \{\{1\},\{0\},\{1\},\{0\},\{1\}\}$.

After that, the algorithm can recompute CS feature (2) and $ISMD_C$ for each pair of clusters according to the Formula (1). Therefore, the CS feature and $ISMD_C$ are the following (the symbol ∪ denotes the pair of clusters):

$CS(x_1 \cup x_2) = \{\{0\},\{1\},\{1\},\{1\},\{0\}\}$,
$CS(x_1 \cup x_3) = \{\{\emptyset\},\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\}\}$,
$CS(x_1 \cup x_4) = \{\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\},\{\emptyset\}\}$,
$CS(x_2 \cup x_3) = \{\{\emptyset\},\{0\},\{\emptyset\},\{1\},\{\emptyset\}\}$,
$CS(x_2 \cup x_4) = \{\{\emptyset\},\{0\},\{1\},\{\emptyset\},\{\emptyset\}\}$,
$CS(x_3 \cup x_4) = \{\{1\},\{0\},\{\emptyset\},\{0\},\{1\}\}$,

$ISMD_C(x_1 \cup x_2) = 0$,

$$ISMD_C(x_1 \cup x_4) = \frac{4}{\sqrt{2}},$$
$$ISMD_C(x_2 \cup x_3) = \frac{3}{2\sqrt{2}},$$
$$ISMD_C(x_2 \cup x_4) = \frac{3}{2\sqrt{2}},$$
$$ISMD_C(x_3 \cup x_4) = \frac{3}{4\sqrt{2}}.$$

The minimum of $ISMD_C$ for each pair of clusters is $ISMD_C(x_1 \cup x_2)$, which is equal to zero. Following Formula (4), the missing value $a_2(x_2)$= * can be replaced as $a_2(x_2) = 1$. After this replacement

$$\{X_1^1, X_2^1, X_3^1\} = \{y_1, y_2, y_3\} = \begin{pmatrix} \{0 & 1 & 1 & 1 & 0\} \\ \{0 & 1 & 1 & 1 & 0\} \\ \{1 & 0 & 0 & * & 1\} \\ \{1 & 0 & 1 & 0 & 1\} \end{pmatrix},$$

where $X_1^1 = X_1^0 \cup X_2^0 = y_1$. The CS feature for y_1 is equal to $CS(y_1)$={{0},{1},{1},{1},{0}}.

The CS feature for clusters y_2 and y_3 remains the same as in the first step, specifically $CS(y_2)$ = {{1},{0},{0},{*},{1}}, and $CS(y_3)$ = {{1},{0},{1},{0},{1}}.

Based Formula (2):

$$CS(y_1 \cup y_2) = \{\{\emptyset\},\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\}\},$$
$$CS(y_1 \cup y_3) = \{\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\},\{\emptyset\}\},$$
$$CS(y_2 \cup y_3) = \{\{1\},\{0\},\{\emptyset\},\{0\},\{1\}\},$$

and $ISMD_C$ then

$$ISMD_C(y_1 \cup y_2) = \frac{4}{\sqrt{3}},$$
$$ISMD_C(y_1 \cup y_3) = \frac{4}{\sqrt{3}},$$
$$ISMD_C(y_2 \cup y_3) = \frac{1}{4\sqrt{2}}.$$

For this step, the minimum of $ISMD_C$ for each pair of clusters is $ISMD_C(y_2 \cup y_3)$. Following Formula (4), the missing value $a_4(y_2)$ = * can be replaced as $a_4(y_2) = 0$. After this replacement, the algorithm will be stopped, because no missing value remains. The final imputed dataset is equal to

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Feng et al. (2011) introduced the theoretical principles of the IMIC method but did not include an objective summary of its advantages and disadvantages. One of the advantages of this method is that the algorithm is easy to use. There is no parameter that needs to be set up, so no additional knowledge or experience with statistical modeling is needed. Another advantage of the method is that it can be used on a dataset with mixed categorical and numerical values.

However, the advantages mentioned above can also be seen as disadvantages. There are not enough possibilities to improve the accuracy of the result. The main problem of the IMIC method is that it is time-consuming. The time complexity is $O(n^3)$ (Murtagh, 1983), where n is the number of data points.

2 THEORETICAL PRINCIPLES OF THE PROPOSED MODIFICATIONS

Formula (1) does not consider the actual number of different categories. Based on the results of Šulc and Řezanková (2014) it is reasonable to try modifying it like this

$$ISMD_C(X_i^r, Y_j^r) = \frac{\sum_k \frac{n_k^2}{n_k^2+2}\, h(s_k(X_i^r, Y_j^r))}{\sqrt{|X_i^r| + |Y_j^r|} \cdot |\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}|}, \tag{5}$$

where:

$$h(s_k(X_i^r, Y_j^r)) = \begin{cases} 1, & s_k(X_i^r, Y_j^r) = \emptyset \\ 0, & s_k(X_i^r, Y_j^r) \neq \emptyset \end{cases}, \tag{6}$$

where the factor $\frac{n_k^2}{n_k^2+2}$ is inspired by the Eskin similarity measure proposed by Eskin et al. (2002), where n_k is the number of categories in the kth variable. If the original algorithm encounters different categories in the cluster in the kth variable, it increases the numerator by one. The possible advantage of our modification is that the $ISMD_C$ can consider the actual number of different categories in the cluster.

The possible problem with this approach is that the situation when the categories are different can occur rarely. In such case, the impact of this improvement can be hardly detected. Given this fact, the equation can be changed as follow:

$$ISMD_C(X_i^r, Y_j^r) = \frac{\sum_k \frac{n_k^2}{n_k^2+2}\, h(s_k(X_i^r, Y_j^r)) + 1)}{\sqrt{|X_i^r| + |Y_j^r|} \cdot (|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}| + 1)}. \tag{7}$$

The difference from previous improvement, the numerator is increasing for each kth variable by factor $\frac{n_k^2}{n_k^2+2}$ and multiply by the number of categories plus one. Adding one to the $h(s_k(X_i^r, Y_j^r))$ ensures that the numerator is always non-zero. Therefore, the factor $\frac{n_k^2}{n_k^2+2}$ is not negligible even if the $h(s_k(X_i^r, Y_j^r))$ equals zero.

3 DATA SOURCE AND APPLICATIONS OF THE PROPOSED MODIFICATIONS

For our experiment, we use the data about students during the 2005–2006 school year, collected by Cortez and Silva (2008). We chose the subset of 395 observations (students who attended the mathematical class) with the following 17 categorical (binary or nominal) variables (of 33 variables overall):

- Sex – student's sex (binary: female or male),

- School – student's school (binary: Gabriel Pereira or Mousinho da Silveira),
- Address – student's home address type (binary: urban or rural),
- Pstatus – parent's cohabitation status (binary: living together or apart),
- Mjob – mother's job (nominal, teacher, health care related, civil services (e.g. administrative or police), at home or other),
- Fjob – father's job (nominal, teacher, health care related, civil services (e.g. administrative or police), at home or other),
- Guardian – student's guardian (nominal: mother, father or other),
- Famsize – family size (binary: less than or equal to 3 or greater than 3),
- Reason – reason to choose this school (nominal: close to home, school reputation, course preference or other),
- Schoolsup – extra educational school support (binary: yes or no),
- Famsup – family educational support (binary: yes or no),
- Activities – extra-curricular activities (binary: yes or no),
- Paidclass – extra paid classes (binary: yes or no),
- Internet – Internet access at home (binary: yes or no),
- Nursery – attended nursery school (binary: yes or no),
- Higher – wants to take higher education (binary: yes or no),
- Romantic – with a romantic relationship (binary: yes or no),
- PassXfail – created variable based on true student's score which shows if students pass the exam or not (binary: yes or no).

The initial dataset is complete without any missing data. In our experiment, the missing values were created in each of the 17 variables separately in five different ratios (5%, 15%, 25%, 35%, and 45%). In each of these configurations, the missing values were created randomly (as MCAR). It is also possible to create the missing values in the whole dataset at once, but there should be no difference due to twenty replications of the experiment. The three methods of the missing imputation were used – the original IMIC algorithm, Modification 1 based on Formula (5), and Modification 2 based on Formula (7). These methods were implemented in the R environment (R Core Team, 2020) and the package RCPP (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2018) was used in crucial parts for better performance.

The algorithm was executed twenty times for each of the five ratios and three versions of the IMIC algorithm for better result stability. In each of these twenty steps, the missing values were generated independently. Therefore, there were 300 runs of the algorithm in summary. After that, the results were averaged based on the specific missing values ratio and the algorithm version.

This setting allows comparing the accuracy of the algorithm based on the specific method and the missing value ratio. For binary variable (with values "0" and "1"), the accuracy can be defined as

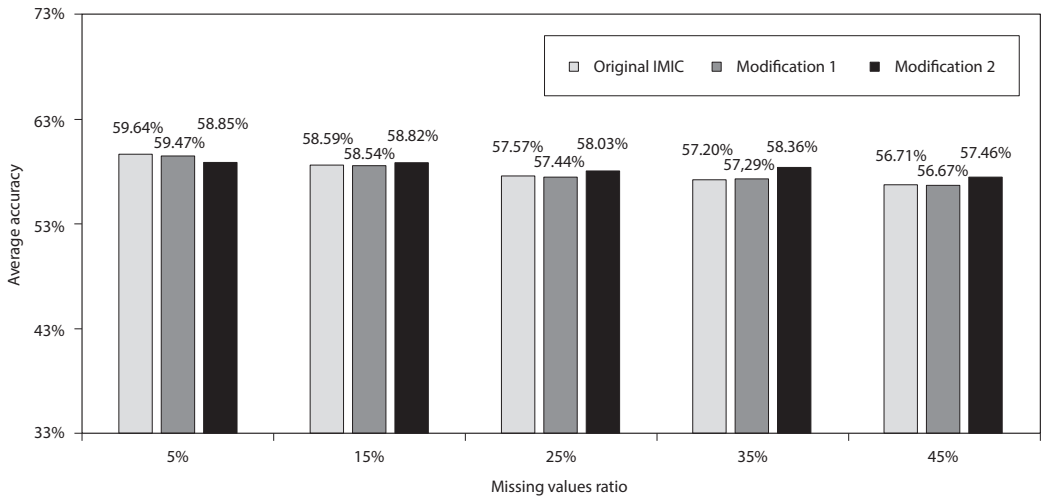
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

where *TP* stands for true positive (the missing value is imputed by "1" correctly), *TN* for true negative (the missing value is imputed by "0" correctly), *FP* for false positive (the missing value is imputed by "1" falsely), and *FN* for false negative (the missing values is imputed by "0" falsely). This formula is defined for binary classification only, but the mean accuracy can be obtained in multiple classification cases. In this paper, the final overall mean accuracy is computed as mean accuracy over all variables and all twenty repetitions.

4 EXPERIMENTAL RESULTS

This section is focused on the simulation evaluations obtained on the dataset collected by Cortez and Silva (2008). The results in Figure 1 show that the Modification 1 works as well as the original IMIC. The Modification 2 works worse when the missing ratio is low, but the mean accuracy improves as the ratio of the missing values grows. The difference in overall average accuracy among these methods is not that large in absolute value, but as presented below, the pairwise t-test shows that the Modification 2 works significantly better than the original IMIC on the dataset used.

Figure 1 Average accuracy for different ratios of missing values (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

As illustrated in Figure 1, when the ratio of the missing values hits 15%, the Modification 2 starts to be slightly better than the other two implementations. Moreover, if the vector of twenty accuracies of original IMIC from each of twenty replications (average over all used variables) is compared to the same vector evaluated using Modification 2, the difference is noticeable. For measuring this difference, the one-sided pairwise t-test was used, see Table 1. When the ratio of the missing values is equal to 35%, the Modification 2 becomes significantly better than the original IMIC algorithm on the dataset used.

Table 1 Comparison of the Modification 2 with the original IMIC algorithm (percent of the missing values and p-values for the t-test)

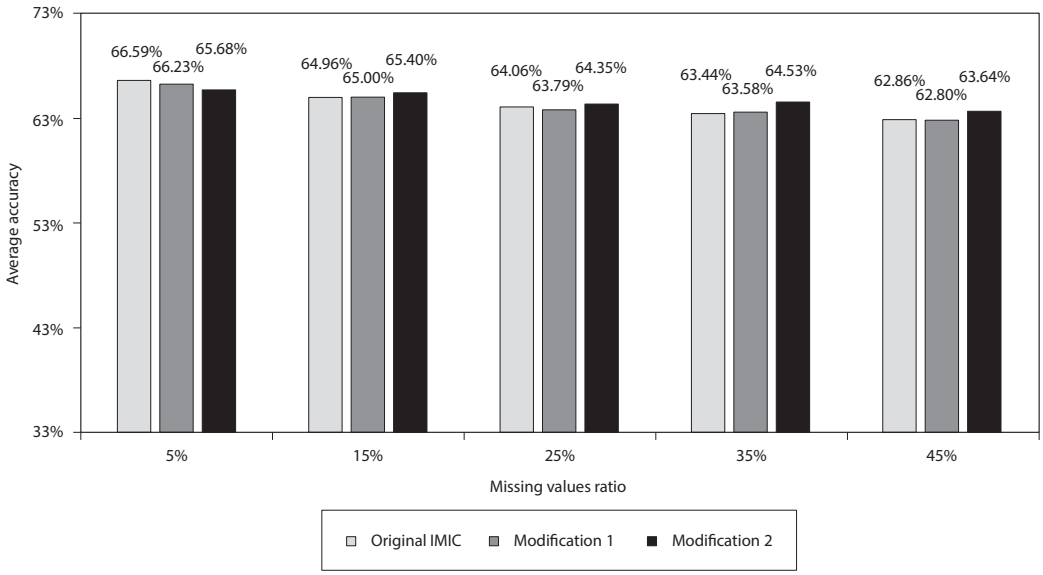
Share of missing values	P-value
5%	0.859
15%	0.266
25%	0.069
35%	< 0.001
45%	0.001

Source: Data collected by Cortez and Silva (2008), own calculation

The dataset can be investigated more deeply. When the variables are split into binary and nominal subsets, the accuracy for binary variables is about 65% (Figure 2) despite the 35% accuracy for nominal variables (Figure 3). The Modification 2 scores better in both cases regardless of the absolute values.

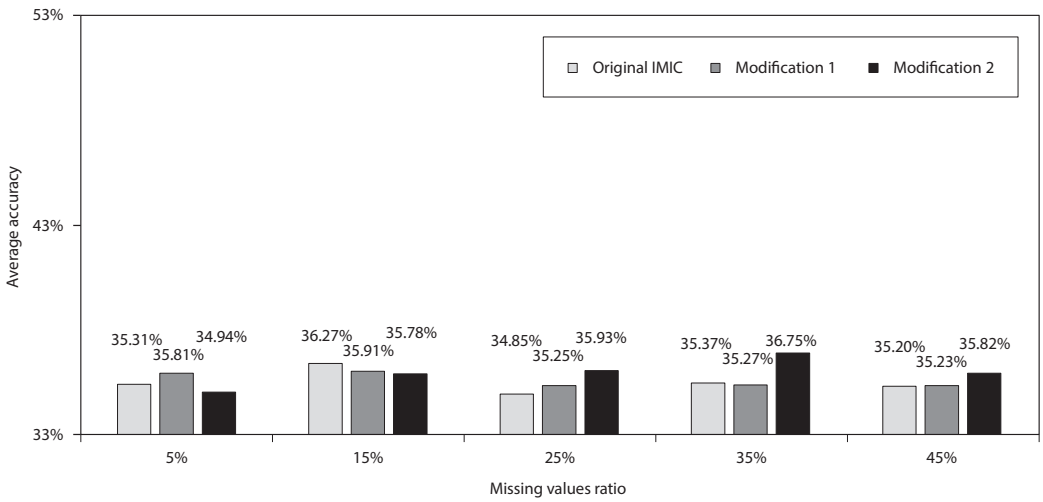
Figure 5 illustrates that the Modification 2 scores better in almost every twenty replications with a 35% missing values ratio compared to Figure 4, which illustrates the same situation with a 5% missing values

Figure 2 Average accuracy for different ratios of missing values (dataset with binary variables)



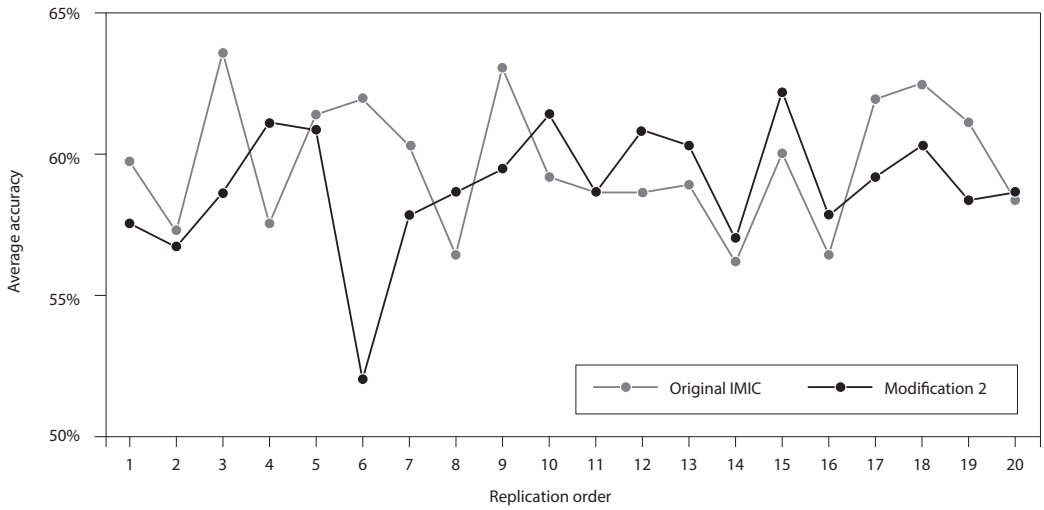
Source: Data collected by Cortez and Silva (2008), own calculation

Figure 3 Average accuracy for different ratios of missing values (dataset with nominal variables)



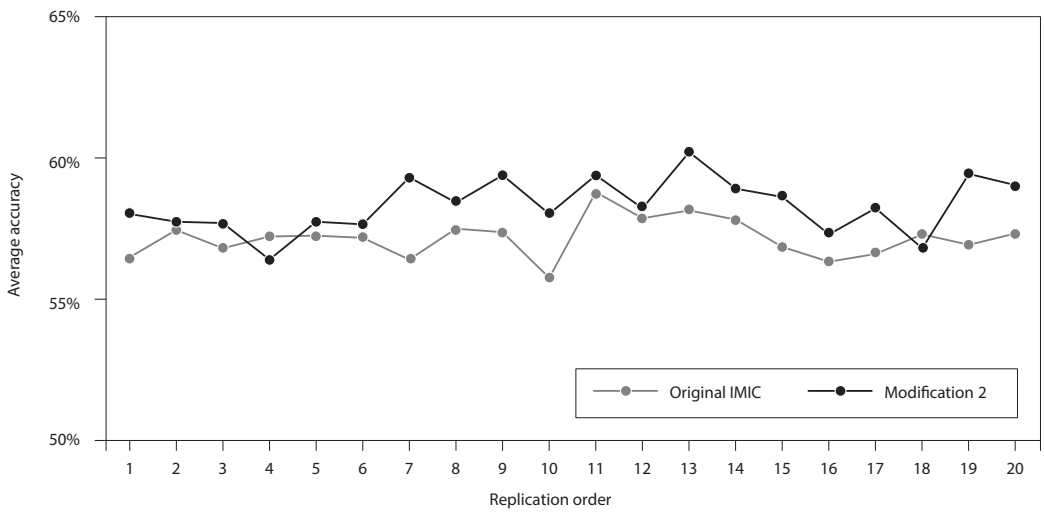
Source: Data collected by Cortez and Silva (2008), own calculation

Figure 4 Average accuracy over variables for 20 replications in 5% missing values ratio setting (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

Figure 5 Average accuracy over variables for 20 replications in 35% missing values ratio setting (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

ratio. It seems that, in the 35% setting, the algorithm is more stable. The coefficient of variation, which is defined as the standard deviation divided by mean, is lower in the 35% missing values ratio setting; concretely, the coefficient of variation for the Modification 2 equals about 0.0161 compared to the 5% missing values ratio setting where the coefficient of variation equals 0.0376.

CONCLUSION

For the purpose of this work, the IMIC algorithm was implemented. This IMIC is easy to use and does not require any additional assumptions on the dataset's properties. It can deal with categorical as well as numerical variables. The main disadvantage lies in time complexity, which is a problem of hierarchical clustering methods in general. Unluckily, this problem makes the simulations very CPU time demanding.

In this paper, two modifications of the IMIC algorithm were proposed and studied on the dataset collected by Cortez and Silva (2008). The first modification, which counts different categories in mismatched observations, was less successful than the second, which considers the overall frequency of categories in each categorical variable in the whole dataset. The differences in accuracy were not too large in absolute values, but the Modification 2 works stably better based on the one-sided pairwise t-test results. These results show the notable difference between accuracy for binary and nominal variables. However, the second modification works better in both cases.

Thanks to the full implementation of the IMIC algorithm, there are many ways for future research. Based on metrics known from hierarchical clustering, the IMIC algorithm can be modified in many different ways. The algorithm, unlike many others, considers the imputed variable itself. Due to this property, it can have some advantages when dealing with MNAR type of missing data. It could be examined in future work. Last but not least, the algorithm should be rewritten for its better efficiency.

ACKNOWLEDGMENTS

This work was supported by the Prague University of Economics and Business under the IGA project No. F4/22/2021.

References

AKANDE, O., LI, F., REITER, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data [online]. *American Statistician*, 71(2): 162–170. <https://doi.org/10.1080/00031305.2016.1277158>.

ALLISON, P. D. (2012). Handling Missing Data by Maximum Likelihood [online]. In: *SAS Global Forum*, paper 312. <https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf>.

AZAR, B. (2002). Finding a Solution for Missing Data [online]. *Monitor on Psychology*, 33(7). <http://www.apa.org/monitor/julaug02/missingdata>.

BARALDI, A. N., ENDERS, C. K. (2010). An Introduction to Modern Missing Data Analyses [online]. *Journal of School Psychology*, 48(1): 5–37. <https://doi.org/10.1016/j.jsp.2009.10.001>.

CIBULKOVÁ, J., NOVÁKOVÁ, L., HORNÍČEK, J. (2021). Imputation Methods for Missing Categorical Data in Cluster Analysis [online]. In: *The 15th International Days of Statistics and Economics*, Prague: Prague University of Economics and Business, 378–388. <https://msed.vse.cz/msed_2021/article/471-Cibulkova-Jana-paper.pdf>.

CORTEZ, P., SILVA, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In: BRITO, A., TEIXEIRA, J. (eds.) *Proceedings of the 5th Annual Future Business Technology Conference*, Porto, 5–12.

EDDELBUETTEL D. (2013). *Seamless R and C++ Integration with Rcpp* [online]. New York: Springer. <https://doi.org/10.1007/978-1-4614-6868-4>.

EDDELBUETTEL D., BALAMUTA J. (2018). Extending extitR with extitC++: a Brief Introduction to extitRcpp [online]. *The American Statistician*, 72(1): 28–36. <https://doi.org/10.1080/00031305.2017.1375990>.

EDDELBUETTEL D., FRANÇOIS R. (2011). Rcpp: Seamless R and C++ Integration [online]. *Journal of Statistical Software*, 40(8): 1–18. <https://doi.org/10.18637/jss.v040.i08>.

DE LEEUW, E. D., HOX, J., HUSMAN, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics*, 19(2): 153–176.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. V. (2002). A Geometric Framework for Unsupervised Anomaly Detection. In: *Applications of Data Mining in Computer Security*, Boston: Springer, 78–100.

FENG, X., WU, S., LIU, Y. (2011). Imputing Missing Values for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical Clustering. In: *International Conference on Knowledge Science, Engineering and Management*, Berlin, Heidelberg: Springer, 414–424.

FERRARI, P. A., ANNONI, P., BARBIERO, A., MANZI, G. (2011). An Imputation Method for Categorical Variables with Application to Nonlinear Principal Component Analysis. *Computational Statistics & Data Analysis*, 55(7): 2410–2420.

MURTAGH, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms [online]. *The Computer Journal*, 26(4): 354–359. <https://doi.org/10.1093/comjnl/26.4.354>

NGUYEN, D. V., YAMADA, K., UNEHARA, M. (2013). Extended Tolerance Relation to Define a New Rough Set Model in Incomplete Information Systems [online]. *Advances in Fuzzy Systems.* <https://doi.org/10.1155/2013/372091>.

PECÁKOVÁ, I. (2014). Problém chybějících dat v dotazníkových šetřeních [online]. *Acta Oeconomica Pragensia*, 22(6): 66–78. <http://www.vse.cz/aop/459>.

PETRÚŠEK, I. (2015). *Analýza chybějících hodnot.* Prague: Sociologický ústav AV ČR.

R CORE TEAM (2020). *A Language and Environment for Statistical Computing* [online]. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.

RUBIN, D. B. (1976). Inference and Missing Data [online]. *Biometrika*, 63(3): 581–592. <https://doi.org/10.1093/biomet/63.3.581>.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Hoboken, New Jersey: John Wiley & Sons.

RUBIN, D. B., LITTLE, R. J. A. (2002). *Statistical Analysis with Missing Data.* 2nd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc., Wiley Series in Probability and Statistics.

SCHAFER, J. L., GRAHAM, J. W. (2002). Missing Data: Our View of the State of the Art [online]. *Psychological Methods*, 7(2): 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>.

SULIS, I., PORCU, M. (2008). *Assessing the Effectiveness of a Stochastic Regression Imputation Method for Ordered Categorical Data.* Working Paper, Centro Ricerche Econoiche Nord Sud.

STAVSETH, M. R., CLAUSEN, T., RØISLIEN, J. (2019). How Handling Missing Data May Impact Conclusions: a Comparison of Six Different Imputation Methods for Categorical Questionnaire Data [online]. *SAGE Open Medicine.* <https://doi.org/10.1177/2050312118822912>.

ŠULC, Z., ŘEZANKOVÁ, H. (2014). Evaluation of recent similarity measures for categorical data [online]. In: *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics*, Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 249–258. <https://doi.org/10.15611/amse.2014.17.27>.

WU, S., FENG, X., HAN, Y. et al. (2012). Missing Categorical Data Imputation Approach Based on Similarity. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2827–2832.