IMPUTATION METHODS FOR MISSING CATEGORICAL DATA IN CLUSTER ANALYSIS

Jana Cibulková – Lucie Nováková – Jaroslav Horníček

Abstract

The demand for analyzing categorical data has increased enormously in recent years. Categorical data became more common in surveys and datasets, resulting in growth in the theory and application of models for categorical data. Handling missing values in a dataset is a problem that frequently arises in statistical practice, and datasets containing categorical data are no exception. Despite this, methods for handling missing values in categorical data analysis have not been readily available. Yet, most of the more sophisticated methods for handling missing values are focused on continuous variables only. Hence data analysts often resort to ad hoc methods of case deletion or very basic imputation methods to transform an incomplete dataset into a complete one. In the paper, we compare several imputation methods from the most basic ones (mode imputation) to the more advanced ones (the MI algorithm), which are rarely used. The methods are briefly described, their pros and cons are pointed out and lastly, they are compared with respect to cluster analysis results using internal and external evaluation criteria (such as the Rand index).

Key words: missing values, cluster analysis, imputation methods, evaluation criteria.

JEL Code: C38, C63

Introduction

Most statistical methods assume a complete dataset to be analyzed. However, some values may be missing in the dataset due to various reasons, and hence they need to be treated. Missing data are a common problem in research studies, especially in questionnaire surveys. A researcher may delete observations with missing values and risk an unwanted loss of information that may potentially bias the output of the analysis; or she/he can choose one of the suitable missing data imputation methods.

Many methods have been suggested for imputing values to missing data. However, existing methods for treating missing categorical data are generally just an adaptation of techniques initially designed for quantitative variables (Ferrari et al., 2011). Also, approaches

used in quantitative data analysis, such as replacement by averages, often cannot be used when dealing with categorical data.

Recently, several methods have been proposed, but they were not sufficiently evaluated, especially in the field of cluster analysis. In (Schafer, 2000), a log-linear model-based multiple imputation approach for missing categorical data was described. However, it is constrained in terms of application due to problems of estimating the higher-order interactions. Another alternative method based on the multinomial distribution and logistic regression was introduced in (Sulis & Porcu, 2008). According to Akande et al. (2017), there are currently three default multiple imputation methods for categorical data. Namely chained equations using generalized linear models, chained equations using classification and regression trees, and a fully Bayesian joint distribution based on Dirichlet process mixture models. Other authors such as Josse and Husson (2016) are focusing on methods in the area of multiple correspondence analysis or factor analysis.

In this paper, we focus on three methods suitable for categorical data and how they may affect the outcomes of cluster analysis. These methods vary in difficulty level from the most basic ones (mode imputation) to the more advanced ones (the MI algorithm). The aim of the paper is to point out the importance of treating missing values in a dataset and demonstrate the effect of various methods of treating missing values (in datasets with categorical data) on the outcomes of cluster analysis.

1 Missing data mechanism

Rubin and Little (2002) classified missing data problems into three categories - MCAR, MAR, and MNAR, based on so-called missing data mechanism and missing data model. Every value in a dataset has some probability of being missing, and the underlying process that governs these probabilities is called a missing data mechanism. The model for the process is called the missing data model.

- If the probability of being missing is the same for all cases, then the *data are said to be missing completely at random* (MCAR). Hence, we assume that the causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing, apart from the apparent loss of information. MCAR data might occur when a respondent simply overlooks a question. While convenient, MCAR is often unrealistic for real-life datasets (van Buuren, 2018).

- If the probability of being missing is the same only within groups defined by the observed data, then the *data are missing at random* (MAR). So, the probability of a value being missing is dependent on some measurable characteristic of the individual but not on the missing value itself. In the context of the survey, if one gender is more likely not to answer a particular question, we may consider this to be MAR data. MAR is a much broader class than MCAR, and it is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption even though it cannot be verified if the data really is MAR (Allison, 2009).
- If neither MCAR nor MAR holds, then we speak of *missing not at random* (MNAR). The likelihood of a variable value being missing is directly related to the value of the variable itself. We assume that the probability of being missing varies for reasons that are unknown to us, for example, they may be considered too private/sensitive to respondents.

2 Methods for treating missing categorical data

If missing data are few and sparse, methods for treating missing data have no substantial impact on the results of further analysis. On the contrary, if missing data are significantly present in the dataset, outcomes of the whole analysis may be greatly influenced by the chosen method to treat missing data. In this section, we introduce four methods suitable for handling datasets with missing categorical data.

2.1 Complete-case analysis

Obviously, the best way to treat missing data is not to have them in the first place (Orchard and Woodbury, 1972, p. 697). The complete-case analysis is the simplest and the oldest method of handling missing data. According to van Buuren (2018) It can be considered the standard approach to missing data.

The complete-cases analysis removes all observations that contain a missing value from a dataset which leads to creating the new reduced dataset. Any further analysis is performed on this reduced dataset. This method assumes MCAR data. If (and only if) data are MCAR, then the reduced dataset is a simple random sample of the original dataset; hence no bias of estimates of mean, variance, or regression coefficients occurs (Rubin & Little, 2002). Another advantage of this method is that it can be used with any data type. On the other hand, unwanted information loss is one of the disadvantages of the compete-case analysis method. When large amounts of

data are missing, the information loss can be so severe that even the power of statistical tests can be compromised (de Leeuw, Hox & Huisman, 2003). Hence, even though this method is the simplest of all, one should consider its disadvantages.

2.2 Mode imputation

The most common imputation – the imputation with the average is not possible in the case of categorical data. For this reason, a technique that replaces a missing value with its most frequent category is a common practice. However, this approach leads to the overrepresentation of the most observed categories. This consequently leads to obvious disadvantages: it underrepresents the variability in the data, and it also completely ignores the correlations between the various components of the data. Hence the outcomes of the analysis might be biased.

2.3 Multiple imputation

The use of the multiple imputation method assumes a probability distribution underlying the data. Based on this probability model, parameter estimates are made using the Bayesian posterior distribution based upon the likelihood function of the proposed model, the observed data, and a prior distribution. The Markov Chain Monte Carlo method of data augmentation is used to get this posterior distribution from which the imputed values for the missing observation are drawn. This imputation process is repeated times to create independent data sets ("imputation phase"). Then "the analytical phase" happens, where the desired analysis is performed on each of the datasets. In the third phase, "the combining phase", the parameter estimates are then combined into a single using a simple arithmetic average (Rubin, 1987).

Schafer (1997) described an imputation approach for categorical variables that was similar to multiple imputation for quantitative data, but this approach can hardly be applied in real-world situations due to its complexity, hence even Schafer (1997, p. 148) suggests the use of the multiple imputation for quantitative data approach instead, with the user rounding the imputed values to fit with the possible values of the variables.

This method can be applied to various data types; it assumes MAR data in order to obtain unbiased estimates and estimates' errors; it is often included in statistical software (van Buuren, 2018). However, it is important to note that this method is not used to obtain one full dataset with no missing observations, unlike previously mentioned methods. Multiple imputation method performs the analysis times in the "analytical phase" and combines the outcomes into one at the very end.

3 Experiment design

We demonstrate differences among methods of treating missing categorical data based on outcomes of hierarchical clustering of a nominal dataset. We analyze a dataset of 16 observations where each observation represents the weather conditions of a given day. The dataset consists of four variables:outlook – three categories (rainy, overcast, sunny)

- temperature – three categories (cool, mild, hot)
- humidity – two categories (high/normal)
- windy – two categories (TRUE/FALSE)

We randomly delete 5%, 15%, and 25% of values in the dataset, and we apply each of the methods from Section 2 on each dataset. Then we proceed to clustering of all datasets.

We use ES similarity measure (Eskin & al., 2002), in average linkage hierarchical clustering process, which is defined as follows: Let us denote the categorical data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, \dots, n$ and $c = 1, \dots, m$; n is the total number of observations; m is the total number of variables. The number of categories of the c-th variable is K_c. Then similarity $ES(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as

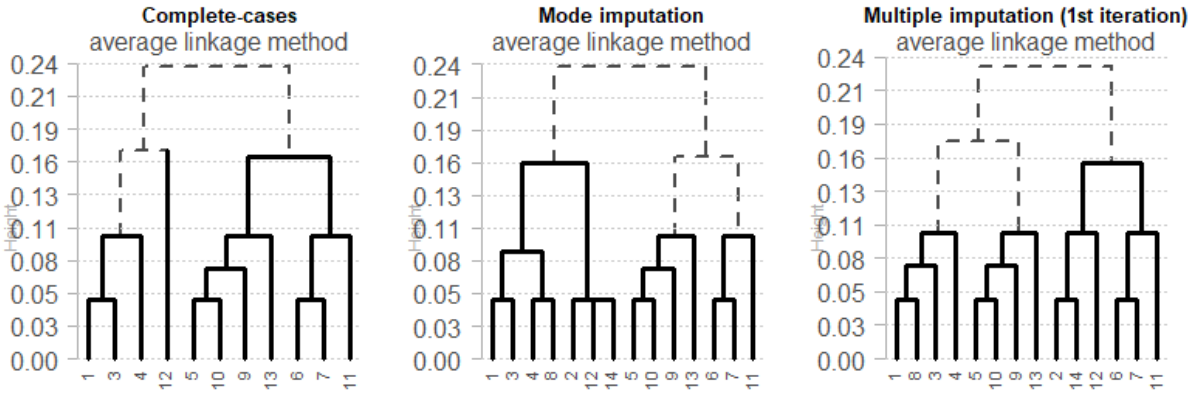
$$ES(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{m}\sum_{c=1}^{m} ES_c(x_{ic}, x_{jc}), \tag{1}$$

where

$$ES_c(x_{ic}, x_{jc}) = \begin{cases} 1; if\ x_{ic} = x_{jc} \\ \dfrac{K_c^2}{K_c^2 + 2}; if\ x_{ic} \neq x_{jc} \end{cases}. \tag{2}$$

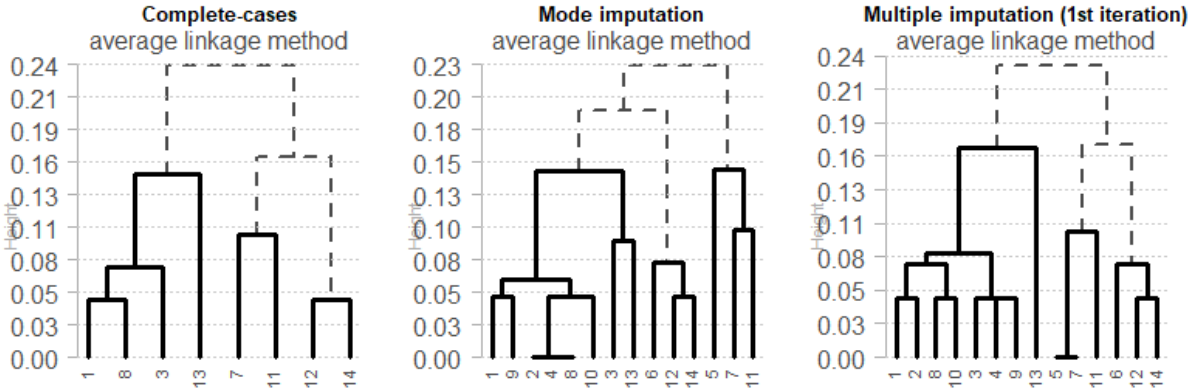
The outcomes of cluster analysis are compared using dendrograms and based on the percentage of identical objects-into-clusters assignments. Triplets of dendrograms corresponding to three different methods of dealing with a given percentage of missing observations are presented in Figures 1 – 3. The percentage gives a value between 0 and 1, where 1 means the two clustering outcomes match identically. The percentage of identical objects-into-clusters assignments is also known as the Rand index (Rand, 1971). For multiple imputation methods, the final assignation of observations into clusters is calculated as the arithmetic mean of all iterations (M=5). It only makes sense to compare mode imputation with multiple imputation. While for 5% missing data is Rand index equal to 1, when 15% of data is missing the Rand index drops to 0.714, and when 25% of data is missing Rand index is equal to 0.725.

Fig. 1: Dendrograms of various methods for 5% missing data



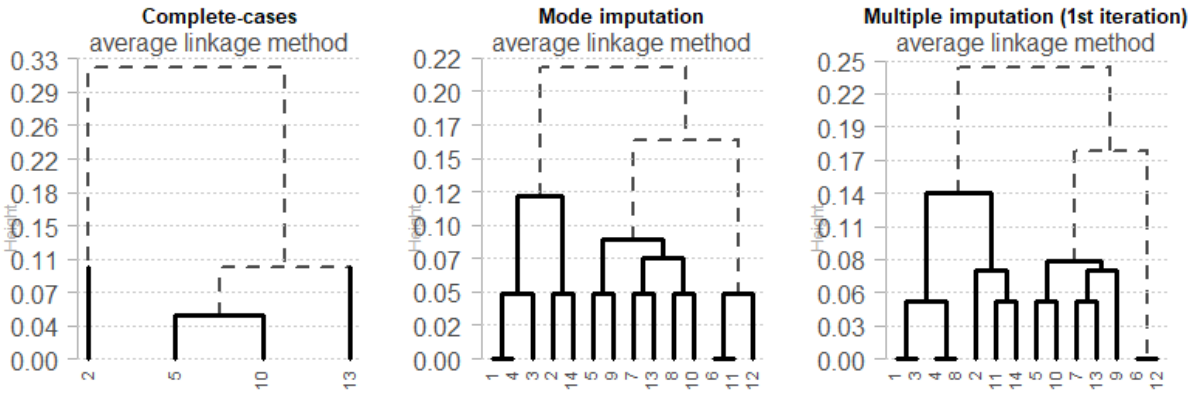
Source: The authors

Fig. 2: Dendrograms of various methods for 15% missing data



Source: The authors

Fig. 3: Dendrograms of various methods for 25% missing data



Source: The authors

Conclusion

Even though the demand for analyzing datasets with categorical data is increasing, statistical methods suitable for categorical data are far behind the methods for quantitative data, in general. The paper aims to provide a brief insight into the problematics of missing categorical data; introduces the current state of knowledge and common practice; points out the gap of quantity and development between methods suitable for quantitative and qualitative data; demonstrates the importance of presented methods on a dataset and observes the effects of dealing with missing observations on the outcomes of cluster analysis. We discussed the pros and cons of several methods, and we also briefly discussed new approaches in the field. Lastly, we demonstrated that outcomes of the clusters analysis obviously differ based on the selected method.

Acknowledgment

This work was supported by the University of Economics, Prague under Grant IGA F4/22/2021.

References

Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. In *American Statistician*, 71(2), 162–170.

Allison, P. D. (2009). Missing Data. In *The SAGE Handbook of Quantitative Methods in Psychology*. London: SAGE Publications Ltd., 72–89. ISBN 978-1-4129-3091-8.

de Leeuw, E. D., Hox, J. & Husman, M. (2003). Prevention and treatment of item nonresponse. In *Journal of Official Statistics*, 19, 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, 6, 78-100.

Ferrari, P. A., Annoni, P., Barbiero, A., & Manzi, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. In C*omputational Statistics & Data Analysis*, 55(7), 2410–2420. https://doi.org/10.1016/j.csda.2011.02.007

Josse, J. & Husson, F. (2016) missMDA: A package for handling missing values in multivariate data analysis. In *Journal of Statistical Software*, 70(1).

Orchard, T. & Woodbury, M. (1972). A missing information principle: Theory and Applications. In *Theory of Statistics*. Berkeley: University of California Press, 1, 697-716. https://doi.org/10.1525/9780520325883-036

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. In *Journal of the American Statistical Association*, 66, 846-850.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

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. ISBN 0-471-18386-5.

Schafer, J. (2000). *Analysis of Incomplete Multivariate Data.* Boca Raton: Chapman & Hall/CRC.

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.

van Buuren, S. (2018). *Flexible Imputation of Missing Data.* 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press. ISBN 978-1138588318.

Contact

Jana Cibulková
University of Economics, Prague
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
jana.cibulkova@vse.cz

Lucie Nováková
University of Economics, Prague

W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
novl15@vse.cz

Jaroslav Horníček
University of Economics, Prague
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
horj31@vse.cz