Małgorzata Misztal

University of Łódź, Faculty of Economics and Sociology,
Department of Statistical Methods, mmisztal@uni.lodz.pl

Comparison of Selected Multiple Imputation Methods for Continuous Variables – Preliminary Simulation Study Results

Abstract: The problem of incomplete data and its implications for drawing valid conclusions from statistical analyses is not related to any particular scientific domain, it arises in economics, sociology, education, behavioural sciences or medicine. Almost all standard statistical methods presume that every object has information on every variable to be included in the analysis and the typical approach to missing data is simply to delete them. However, this leads to ineffective and biased analysis results and is not recommended in the literature. The state of the art technique for handling missing data is multiple imputation. In the paper, some selected multiple imputation methods were taken into account. Special attention was paid to using principal components analysis (PCA) as an imputation method. The goal of the study was to assess the quality of PCA-based imputations as compared to two other multiple imputation techniques: multivariate imputation by chained equations (MICE) and missForest. The comparison was made by artificially simulating different proportions (10–50%) and mechanisms of missing data using 10 complete data sets from the UCI repository of machine learning databases. Then, missing values were imputed with the use of MICE, missForest and the PCA-based method (MIPCA). The normalised root mean square error (NRMSE) was calculated as a measure of imputation accuracy. On the basis of the conducted analyses, missForest can be recommended as a multiple imputation method providing the lowest rates of imputation errors for all types of missingness. PCA-based imputation does not perform well in terms of accuracy.

Keywords: incomplete data, multiple imputation, principal component analysis, missForest

JEL: C18, C80, C38

1. Introduction

Data sets with missing values are quite common in practical applications of statistical methods and, as Allison (2002: 1) points out, "sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data". The occurrence of missing data is not related to any particular scientific domain, it arises in economic, social, educational, behavioural or medical research. It is a problem because almost all standard statistical methods presume that every object has information on every variable to be included in the analysis.

Although, as Orchard and Woodbury (1972: 697) remark: "obviously the best way to treat missing data is not to have them", this way cannot be used in practice and there is a rather strong need for investigating methods dealing with incomplete data.

The typical approach to missing data is to delete them. That is usually the default for statistical packages. This strategy is commonly known as *complete case analysis*. According to van Buuren (2012: 5): "The inclination to delete the missing data is understandable. Apart from the technical difficulties imposed by the missing data, the occurrence of missing data has long been considered a sign of sloppy research. [...] Publication chances are likely to improve if there is no hint of missingness".

To understand why removing objects with missing values from the data set is not the recommended way to solve the problem of missing data occurrence, it is important to distinguish three missing data mechanisms (MDM; Little, Rubin 2002): *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) and *Missing Not at Random* (MNAR). If X is the ($n \times p$) matrix of complete data which is not fully observed, one can divide it into the observed part, denoted by X_{obs}, and the missing part, denoted by X_{mis}. Then:

1) MCAR means that the probability that an item of information is missing does not depend on X_{mis} or on X_{obs};
2) MAR means that the probability that an item of information is missing does not depend on X_{mis}, but may depend on X_{obs};
3) MNAR means that the probability that an item of information is missing does depend on X_{mis}.

With regard to the missing data mechanisms, Enders (2010: 39) emphasises that the standard deletion methods can be used only if the MCAR assumption is met. However, even in that case, eliminating data can lead to serious biases, especially if the amount of missing values is large. If the data are not MCAR, complete case analysis produces biased estimates of means, regression coefficients and correlations (van Buuren, 2012: 8).

It is therefore necessary to search for methods replacing missing data by some plausible values. Some of these methods are briefly described in the paper. Spe-

cial attention is paid to using principal components analysis (PCA) as an imputation method. PCA-based methods are designed for structured data with groups of variables and groups of objects (Josse, 2016: 3). The goal of the study is to assess the quality of PCA-based imputations as compared to some state of the art imputation methods.

2. Imputation methods

The basic idea of imputation is to replace missing values with some reasonable values, based on other available information, and then to do the analysis as if there were no missing data. There are many different ways to fill in missing values. Under the assumption of the MCAR or MAR mechanism, one can use, among others, mean or mode imputation, conditional mean imputation (i.e. regression imputation), stochastic regression imputation, hot deck imputation, substitution, cold deck imputation, the maximum likelihood (ML) method, the EM algorithm, predictive mean matching, k-NN imputation, etc. The MNAR mechanism requires a different and more complex approach, i.e. selection models or pattern-mixture models (see details in Allison, 2002; Little, Rubin, 2002; Enders, 2010; van Buuren, 2012).

To substitute for missing data, single or multiple imputation methods can be applied. In single imputation, missing values are filled in once. In multiple imputation, missing values are filled in m times, statistical analyses are performed on each of the m imputed data sets and the results from the m analyses are combined into one result. Multiple imputation is recommended as an appropriate way of handling incomplete data since it takes into account the uncertainty in the imputations.

Taking into account only continuous data, missing values can be completed with several multiple imputations methods. These are, among others, joint modelling based on the multivariate normal model (Schafer, 1997), multivariate imputation by chained equations (van Buuren, 2007; van Buuren, Groothuis-Oudshoorn, 2011) and missForest (Stekhoven, Bühlmann, 2012). It is also possible, but less popular, to impute missing continuous data with principal components analysis (Josse, Pagès, Husson, 2011; Josse, 2016).

Principal components analysis (Pearson, 1901; Hotelling, 1933) is one of the most popular statistical methods for exploring and analysing multivariate data. It transforms the original variables into new ones (the principal components, defined as linear combinations of the original variables) that are uncorrelated and account for decreasing proportions of the variance in the data. Classical PCA requires complete data sets. Several algorithms have been proposed to make it possible to perform PCA in the presence of incomplete data (for details and theoretical basics see e.g.: Ilin, Raiko, 2010; Josse, Pagès, Husson, 2011; Josse, Husson, 2012; Josse, 2016; Audigier, Husson, Josse,

2016). The primary goal of these methods is to estimate the PCA parameters (scores and loadings) and obtain the associated graphical representations despite missing values. However, as Josse and Husson (2016: 23) claim: "imputation is done during the running of the algorithm and consequently these methods can be used to impute data. Even if at first this 'imputation' may be seen as an aside to these methods, it is in fact very valuable and indeed, the quality of imputation is usually high. This can be explained by the fact that imputation is based on the scores and loadings and thus takes into account similarities between individuals as well as relationships between variables". Two algorithms for multiple imputation via PCA models, i.e. multiple imputation using a parametric bootstrap (Josse, Husson, 2012) and multiple imputation using a Bayesian treatment of the PCA model (Audigier, Husson, Josse, 2016), are implemented in the R environment via the `missMDA` package (Josse, Husson, 2016).

The now classical joint modelling, proposed by Schafer (1997), entails specifying a multivariate normal distribution for the missing data and drawing imputation from their conditional distributions by Markov Chain Monte Carlo (MCMC) techniques. According to van Buuren and Groothuis-Oudshoorn (2011: 2): "this methodology is attractive if the multivariate distribution is a reasonable description of the data".

The methods of multiple imputation currently regarded as the most modern and highly recommended are multiple imputation by chained equations and missForest.

Multivariate Imputation by Chained Equations (MICE) or Fully Conditional Specification (FCS; van Buuren, 2007; van Buuren, Groothuis-Oudshoorn, 2011), known also as sequential regression imputation (Enders, 2010: 275), is based on the iterative process that involves specifying a conditional distribution for each incomplete variable. There is no need to explicitly assume any particular multivariate distribution, it is enough if it can be assumed that one such distribution exists and draws can be generated from it with the use of the Gibbs sampler. The imputed values can be either the predicted values sampled from the posterior distribution of the incomplete variable or obtained using predictive mean matching as the observed value from the complete case with the closest predicted value to the incomplete case (Yu, Burton, Rivero-Arias, 2007: 244).

The missForest method, proposed by Stekhoven and Bühlmann (2012), is an iterative imputation technique based on the Breiman's Random Forests algorithm (Breiman, 2001). A random forest, trained on the observed values of a data matrix, is used to predict the missing values. The advantage of this method is taking into account complex interactions and non-linear relations among variables (see e.g.: Stekhoven, Bühlmann, 2012; Misztal, 2013).

The simulation studies confirm that both methods (i.e. MICE and missForest) perform well and can produce unbiased parameter estimates and standard errors (see e.g.: Shah et al., 2014; Tang, Ishwaran, 2017; Wulff, Ejlskov, 2017). Both of these methods are useful when it is not possible to determine the suitable multivariate distribution.

3. Assumptions of the experiment

As stated above, the objective of the study is to assess the quality of the PCA-based imputations as compared to some other imputation methods. Since the primary goal of the algorithm proposed by Josse, Pagès and Husson (2011) is to perform PCA despite missing values and not to impute missing values per se, it can therefore be interesting to investigate the accuracy of the imputations obtained for the purpose of ascertaining whether the resulting complete data set can be useful for other analyses.

In the simulation study, the results from multiple imputation with the use of the parametric bootstrap PCA (MIPCA) are taken into account and compared with the results from multivariate imputation by chained equations (MICE) and miss-Forest. The benchmark choice is motivated mainly by the impossibility to specify the joint multivariate distribution of the data used in the experiments.

In order to compare all the imputation methods, 10 complete data sets from the UCI repository of machine learning databases (Blake, Keogh, Merz, 1988) and from the author's own research (AR) were selected. A short description of all the data sets is presented in Table 1.

Since PCA-based methods are included in the analysis, it is therefore interesting to look at the values of Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. The KMO values less than 0.50 suggest that PCA probably will not be very useful for structure detection, however, it is not clear whether this will affect imputation accuracy.

Missing data were introduced into each dataset using the function `ampute()` from the `mice` package (v. 2.30, 18.02.2017) and assuming three missing data mechanisms: MCAR, MAR, and MNAR. Five levels of proportion of missing values were considered: 10%, 20%, 30%, 40%, 50%.

Table 1. Short description of data sets used in simulation experiments

Name	Source	Number of cases	Number of variables	KMO
User Knowledge Modelling Data Set	UCI	403	5	0.480
Credits	AR	100	6	0.359
Seeds	UCI	210	7	0.788
AF	AR	300	8	0.590
Glass Identification Data Set	UCI	214	9	0.131
Wine Quality (red)	UCI	1599	11	0.432
Wine RecognitionData	UCI	178	13	0.779
Facebook Performance Metrics	UCI	500	14	0.623
Leaf Data Set	UCI	340	14	0.636
Wisconsin Diagnostic Breast Cancer	UCI	569	30	0.832

Source: author's elaboration

Missing values were imputed with the use of predictive mean matching (PMM; via the `mice` package; with the number of multiple imputation $m = 5$), missForest (mF; via the `missForest` package with 100 trees) and MIPCA with two variants: $\text{dim}^1 = 2$ and $\text{dim} = p - 1$, where p is the number of variables (via the `missMDA` package). RMSE (*root mean square error*) was used as a measure of imputation accuracy. The final RMSE was averaged over the 1000 repetitions. To compare the results for different datasets normalised RMSE (NRMSE) was calculated as RMSE divided by the mean value of the measurements in the original complete data set. Lower values indicate better imputation accuracy.

4. Results

NRMSE values for each analysed data set are presented in Figures 1–10, considering three missing data mechanisms, five levels of missing values fraction and four imputation methods. The lower the NRMSE value the better imputation accuracy.

The lowest NRMSEs (i.e. the highest quality of imputation) were obtained for two data sets: Glass Identification and Seeds. Both these data sets consist of slightly more than 200 observations and are characterised by a relatively small number of variables.

In the case of the Glass Identification data set, the four imputation methods under consideration lead to noticeably different results. The best method of imputation was MIPCA with $p - 1$ (i.e. 8) principal components, the worst results were also observed for MIPCA but with only the first two components taken into account.

The results for the Seeds data set were similar due to the NRMSE value for missForest, MIPCA ($p - 1$) and MICE (PMM) and slightly worse for MIPCA–2.

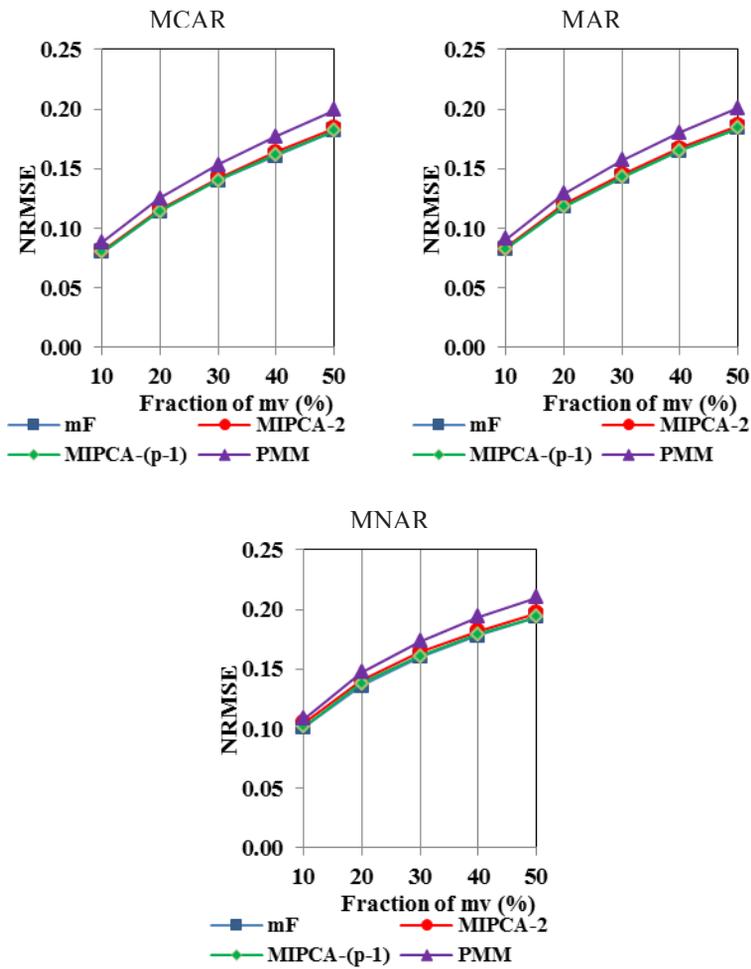[1] The number of principal components taken into account.

Figure 1. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the User Knowledge Modelling Data Set

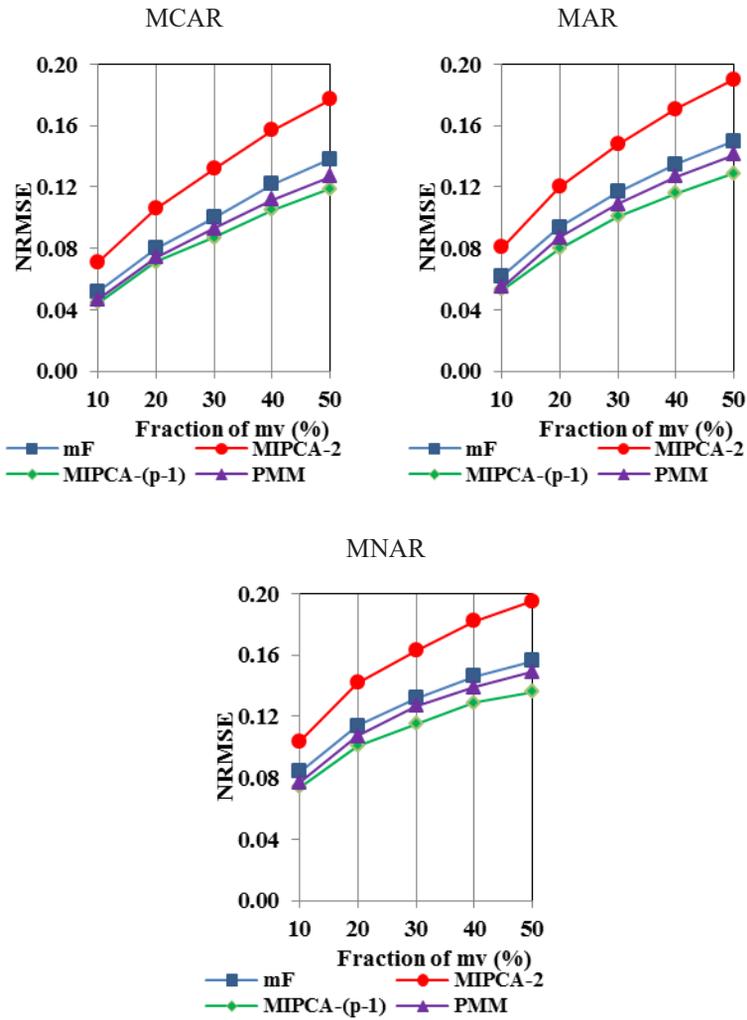Source: own calculations

MCAR

MAR

MNAR

Figure 2. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Credits Data Set
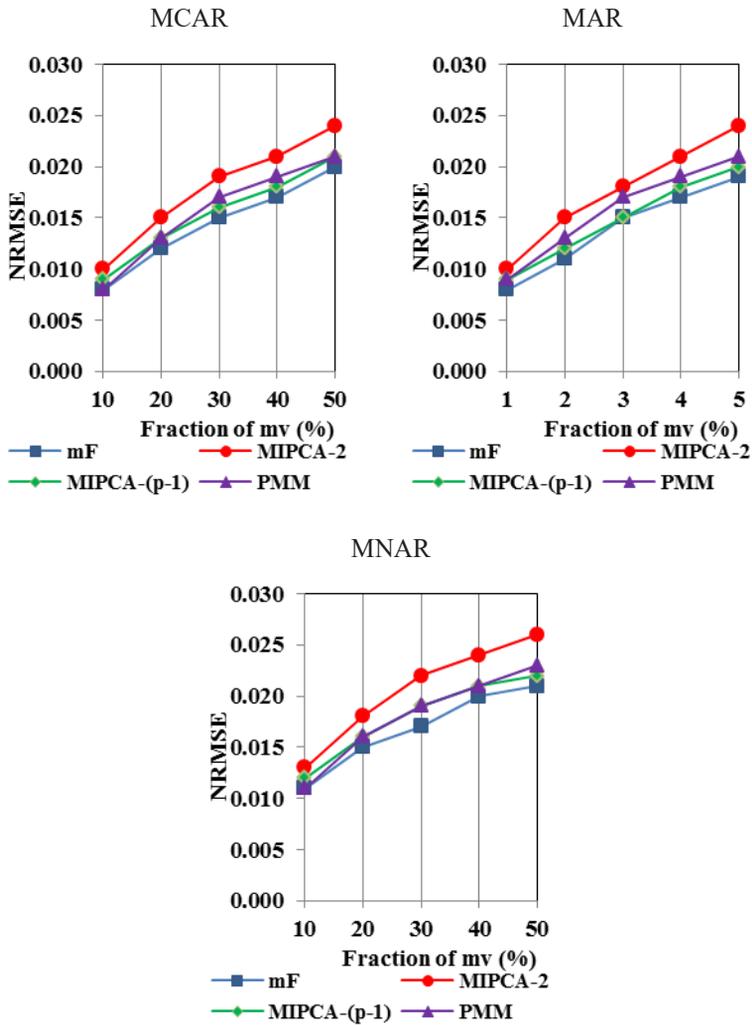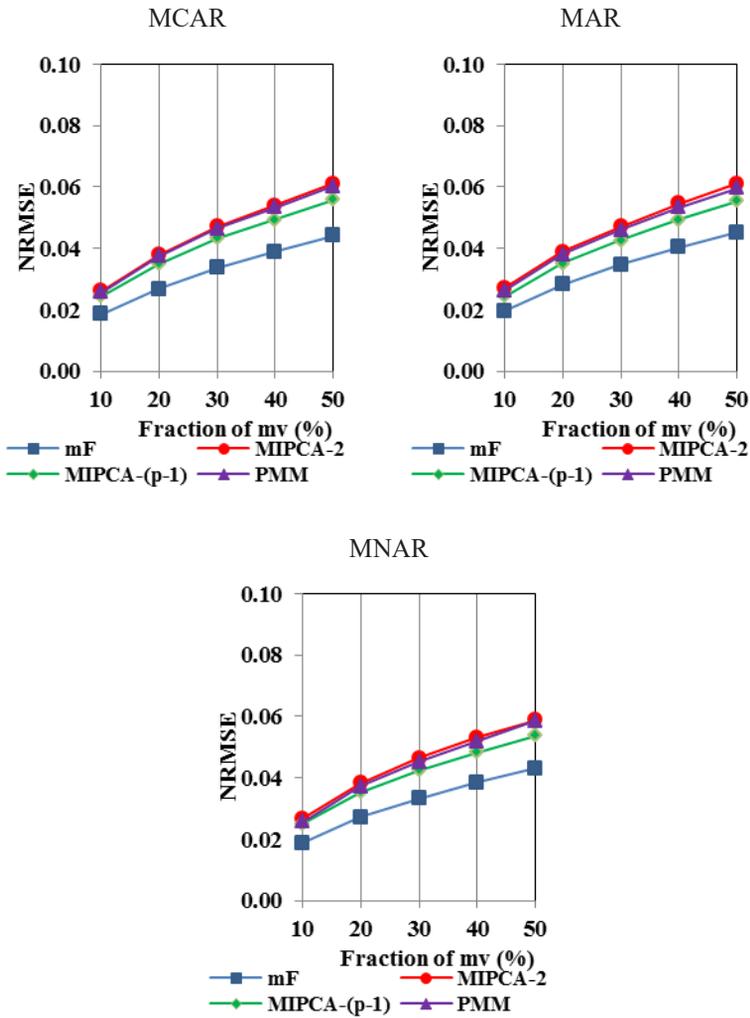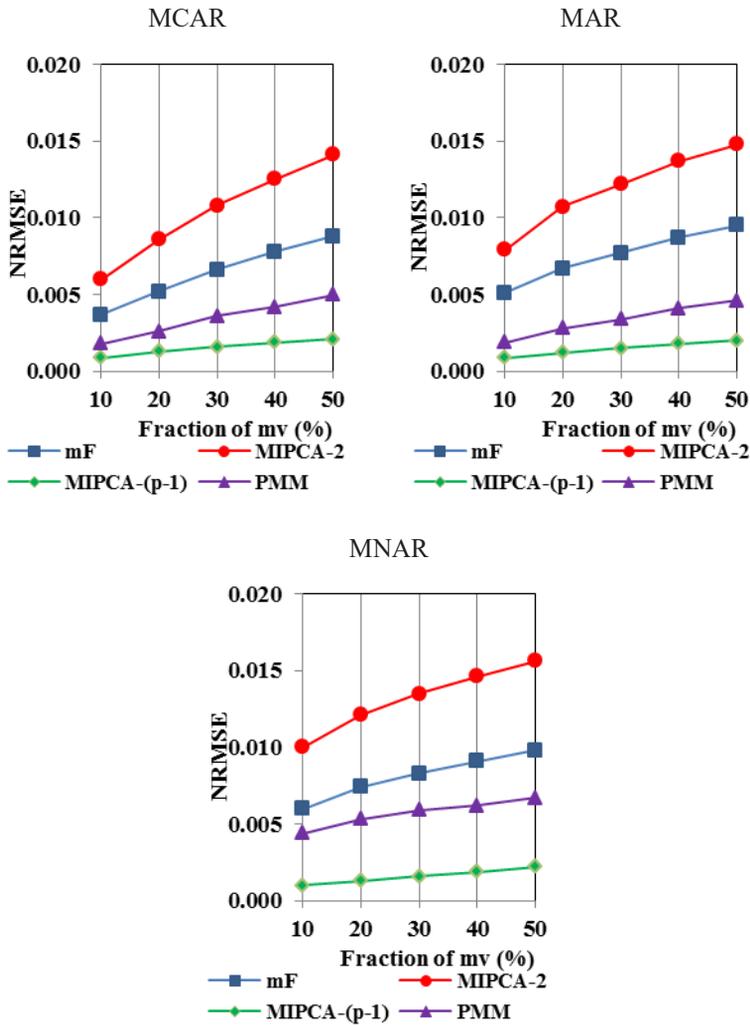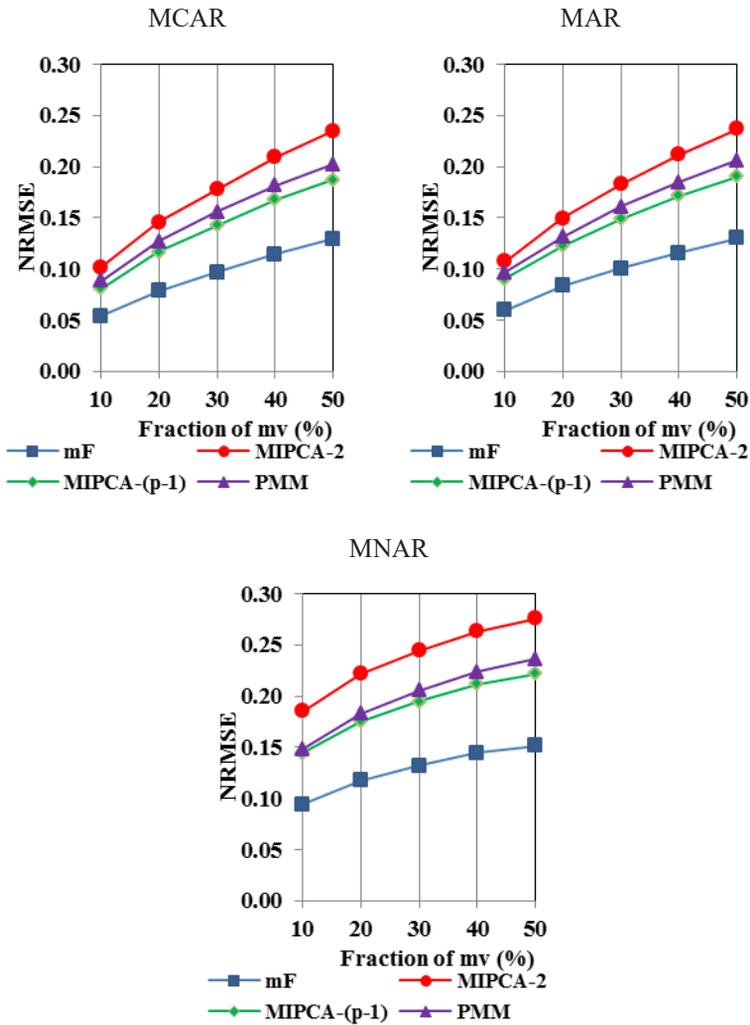
Source: own calculations

MCAR

MAR

MNAR

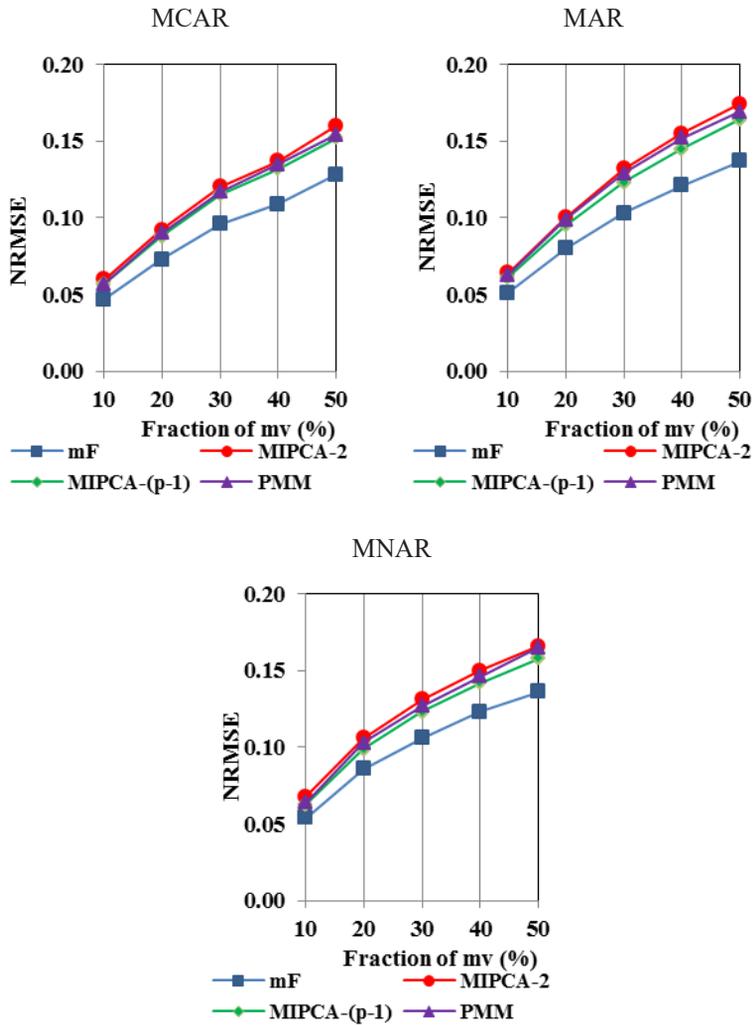Figure 3. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Seeds Data Set

Source: own calculations

MCAR

MAR

MNAR

Figure 4. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the AF Data Set

Source: own calculations

MCAR

MAR

MNAR

Figure 5. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Glass Identification Data Set

Source: own calculations

MCAR

MAR

MNAR

Figure 6. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Wine Quality (red) Data Set

Source: own calculations

MCAR

MAR

MNAR

Figure 7. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Wine Recognition Data Set
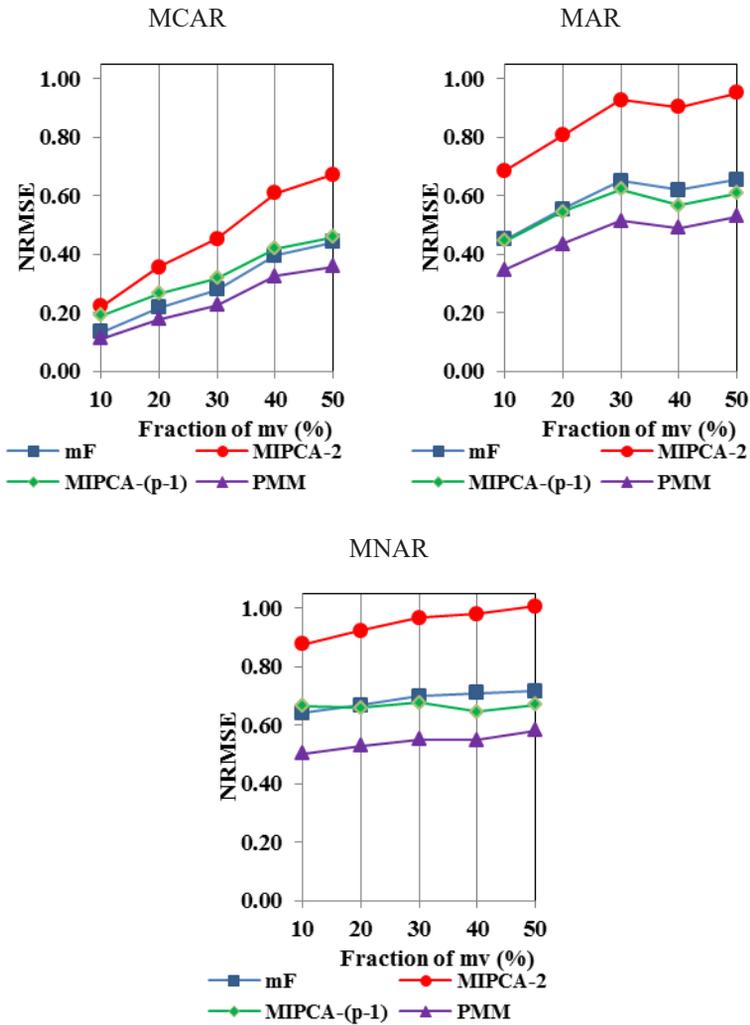
Source: own calculations

Figure 8. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Facebook Performance Metrics Data Set
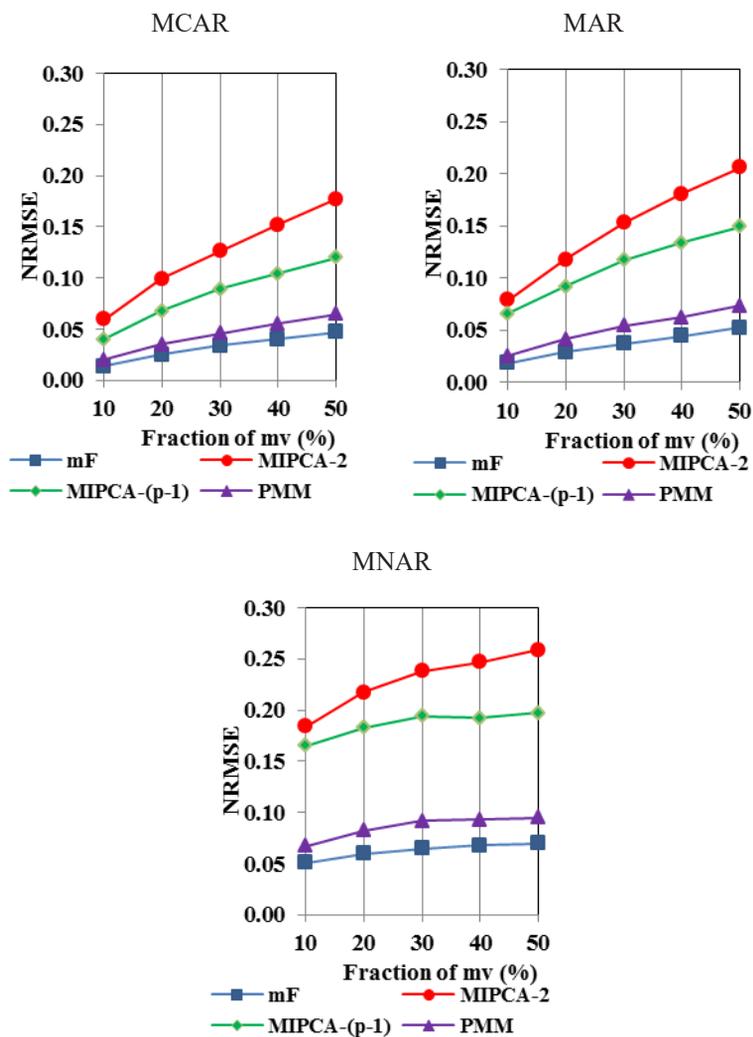
Source: own calculations

Figure 9. NRMSE at increasing missingness levels for different imputations methods assuming
different missing data mechanisms for the Leaf Data Set
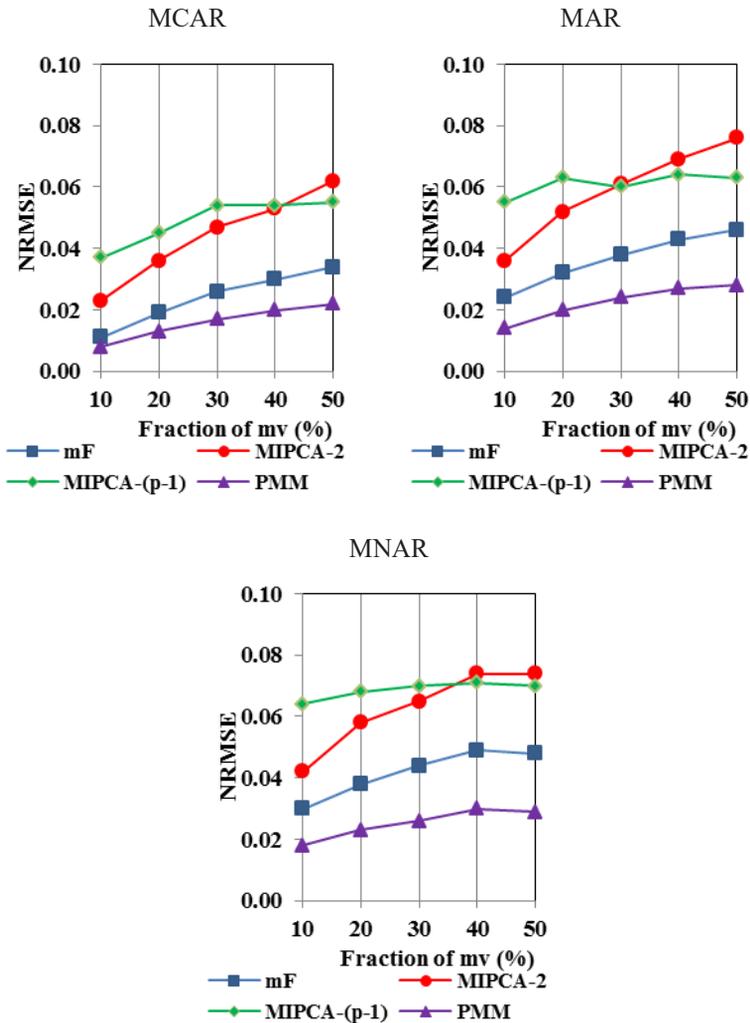
Source: own calculations

Figure 10. NRMSE at increasing missingness levels for different imputations methods assuming different missing data mechanisms for the Wisconsin Diagnostic Breast Cancer Data Set

Source: own calculations

The highest NRMSE values (i.e. the most inaccurate imputation results) were achieved for the Facebook Performance Metrics data set. The MIPCA method with two principal components showed definitely the worst results. The lowest error rate was observed for the predictive mean matching imputation. MIPCA with ($p - 1$) principal components and missForest produced similar results. For that dataset, all the methods used to impute missing values lead to significantly worse results under the MAR and MNAR missing data mechanisms.

Examining all the four imputation methods, it can be observed that MIPCA with two principal components was the least effective imputation technique when compared to the other imputation methods for most of the analysed data sets. For only three data sets (User Knowledge Modelling, AF, Wine Recognition), the results of imputation obtained with this method were similar to those obtained with the other methods. The use of missForest method gave the best imputation results in terms of accuracy for 6 out of 10 data sets. The behaviour of predictive mean matching was not consistent from one dataset to another.

On the basis of the obtained results, it is not possible to explicitly determine the influence of the KMO measure value on imputation accuracy by means of the analysed imputation methods, in particular the PCA-based ones. The MIPCA method with $(p - 1)$ principal components was the most effective in terms of imputation accuracy for the Glass Identification data set, i.e. the one with the lowest KMO, and the most ineffective for the WDBC data set, for which the KMO measure was the highest.

The average performance of the imputation algorithms was also assessed globally on the basis of the NRMSE values for the 10 analysed datasets, considering three missing data mechanisms and five levels of missing data proportions. The results are summarised in Figures 11–13 (medians with interquartile ranges are presented).
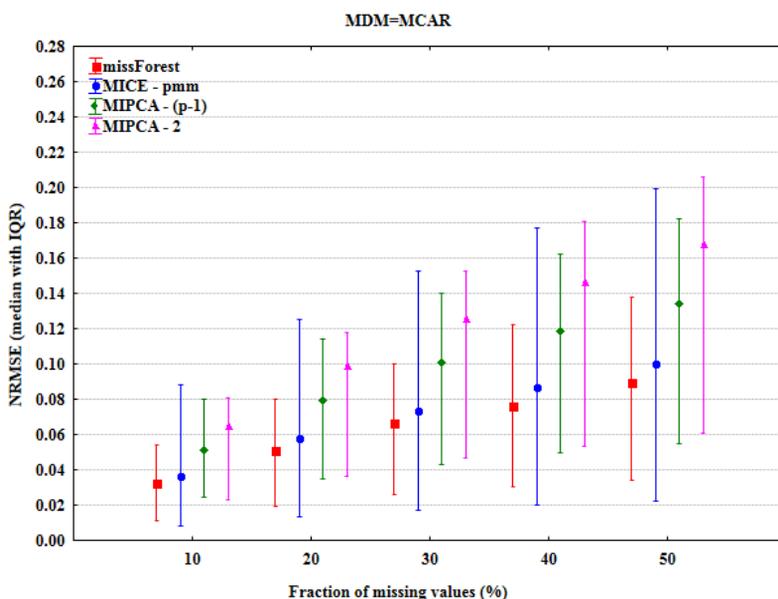

Figure 11. Average performance of the analysed imputation methods at increasing missingness levels under the MCAR missing data mechanism
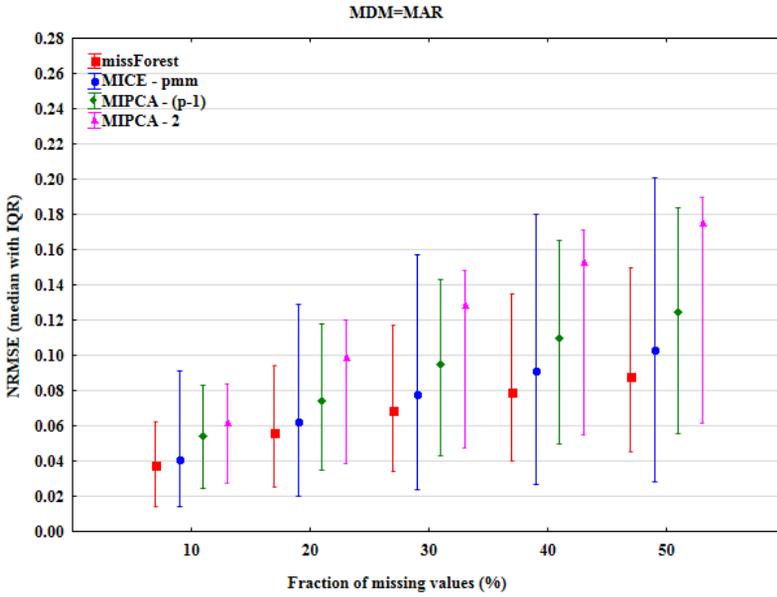
Source: own calculations

Figure 12. Average performance of the analysed imputation methods at increasing missingness
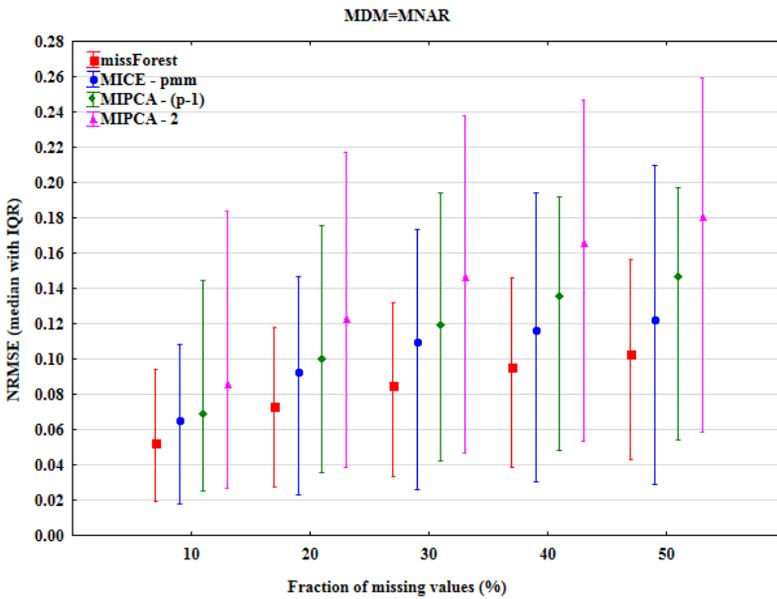levels under the MAR missing data mechanism

Source: own calculations

Figure 13. Average performance of the analysed imputation methods at increasing missingness
levels under the MNAR missing data mechanism

Source: own calculations

The missForest imputation outperforms (on average) all the other methods in terms of imputation accuracy, regardless of the missing data mechanism and the amounts of missing values. The variability of the results for missForest is the lowest when compared to the other methods. The errors obtained from MIPCA imputation with only 2 dimensions are the biggest compared to all the other methods. Taking into account $(p-1)$ dimensions in MIPCA imputation improves the results obtained (decreases NRMSE).

The comparison of the selected imputation methods is also shown in Figures 14–16, where the increase in NRMSE (in %) for all the methods is presented compared to the missForest method (i.e. the one for which the most accurate results were obtained).

The increase in NRMSE is the lowest for predictive mean matching and the highest for MIPCA with 2 dimensions compared to missForest, regardless of the missing data mechanism and the missing data fraction. Taking into account more than the two dimensions in MIPCA, it is possible to reduce the NRMSE value by up to half.
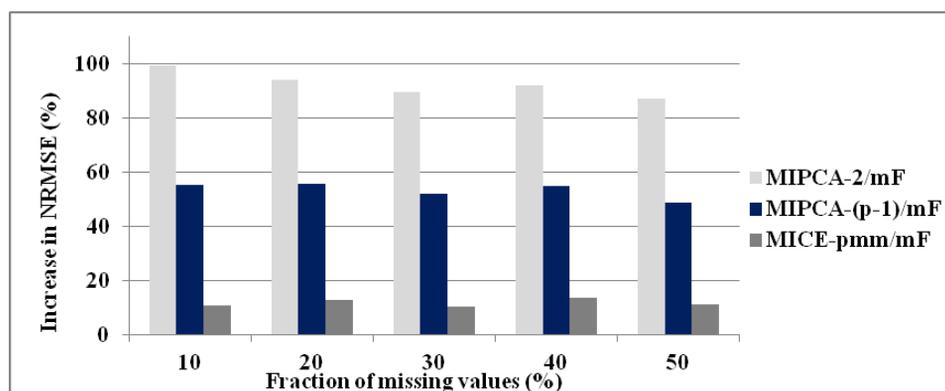

Figure 14. Increase in NRMSE for MIPCA and MICE imputations compared to missForest results under the MCAR missing data mechanism
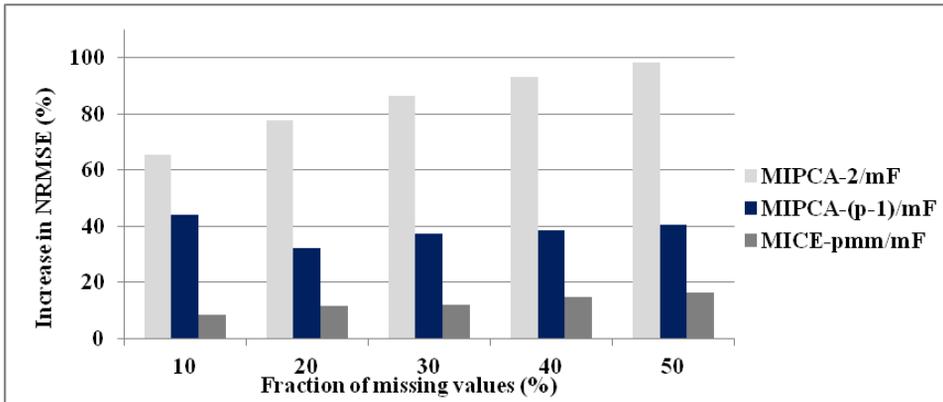
Source: own calculations

Figure 15. Increase in NRMSE for MIPCA and MICE imputations compared to missForest results under the MAR missing data mechanism
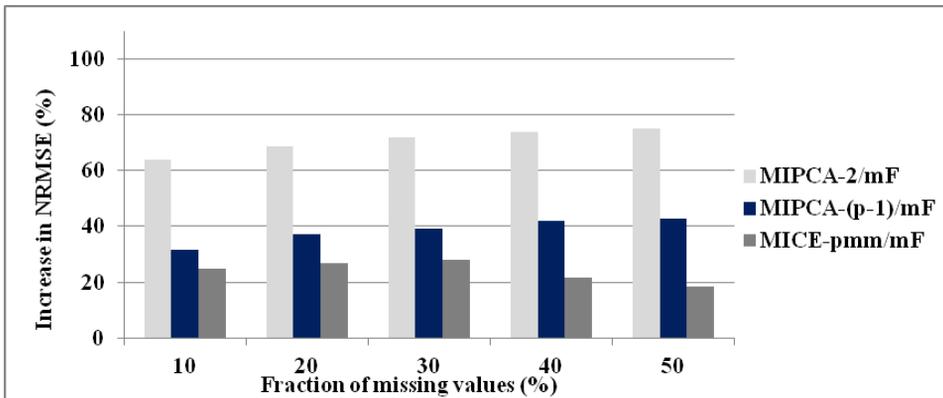
Source: own calculations

Figure 16. Increase in NRMSE for MIPCA and MICE imputations compared to missForest results under the MNAR missing data mechanism

Source: own calculations

To assess the relationships between the missing data mechanisms, fractions of missing values, the methods used to fill in the missing data and the NRMSE value, a regression tree (CART) was applied. The results are presented in Figure 17.
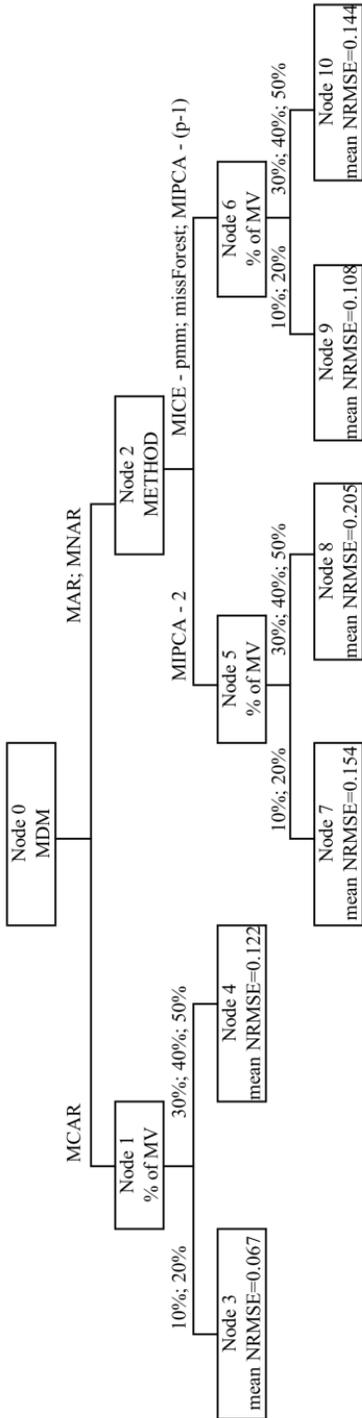
Figure 17. Relationships between the missing data mechanisms, imputation methods, fractions of missing values and the NRMSE values

Source: own calculations

Under the assumption of the MCAR missing data mechanism, imputation quality shall be affected by the amount of missing values in the data set rather than by the imputation method. If the percentage of missing data does not exceed 20%, the lowest NRMSE is observed.

If the MCAR assumption is not met, imputation accuracy shall be influenced by both the choice of the imputation method and the percentage of missing data in the dataset. Handling missing data using the MIPCA method with 2 dimensions provides the worst results (i.e. the highest, on average, NRMSE values are obtained). The other imputation methods are effective (in terms of accuracy) when the fraction of missing values is no more than 20%.

5. Conclusions

Simple and frequently used imputation techniques, such as complete case analysis and overall mean imputation, lead to inefficient analyses results and biased estimates. The methods for handling missing data, recommended in the literature, include, among others, maximum likelihood and multiple imputation (Enders, 2010; Newman, 2014). In the paper, three multiple imputation methods are presented.

When assessing the imputation accuracy measured by the NRMSE value, it is important to consider the missing data mechanism that influences the imputation quality. Therefore, in practical applications, the hypothesis that the missingness is of the MCAR type should be verified with the use of Little's test. The advantages and disadvantages of that test are discussed by Enders (2010: 19–21).

If data are missing completely at random and the amounts of missing values are relatively small (up to 10%), all the imputation methods taken into account perform (on average) in a similar way. Differences in the NRMSE values can be observed with a higher percentage of missing data (30% and more). Lower rates of errors are then obtained for missForest and predictive mean matching imputation compared to MIPCA.

If the missing data mechanism is not MCAR, it shall be MAR or MNAR. Newman (2014: 377) points out: "[…] missing data are almost never missing completely randomly (MCAR). As such, most missing data fall on a continuum between one extreme – where the systematic missingness pattern depends entirely on the observed data (pure MAR), and the other extreme – where the systematic missingness pattern depends entirely on the missing data (pure MNAR). In typical scenarios, systematic missingness depends in part on the observed data (MAR) and in part on the missing data (MNAR), to varying degrees". Even if the MAR assumption is not fully met in practice, multiple imputation techniques (based on this assumption) give better results than some other available and commonly used methods.

On the basis of the conducted analyses, missForest can be recommended as a multiple imputation method providing the lowest rates of imputation errors for both MAR and MNAR missing data mechanisms.

PCA-based imputation does not perform well in terms of accuracy, especially when only the first two dimensions are included in the analysis. In PCA, all the principal components are linear combinations of the original variables and in total account for 100% of the variance of the observations. Taking into account $(p - 1)$ dimensions in MIPCA improves the results (i.e. decreases NRMSE) since almost all available original information is used in the imputation procedure.

Although the results of this study do not support the conclusion that the use of PCA-based imputation method leads to more accurate results than commonly recommended missForest or MICE, this method cannot be regarded as useless. All the results presented in this paper should be viewed as an initial step to more complex analysis of the MIPCA method.

Further research will focus on the use of the Bayesian treatment of the PCA model and its applicability in a case where the number of variables in the data set exceeds the number of observations and these variables are highly correlated.

References

Allison P. D. (2002), *Missing data*, Series: Quantitative Applications in the Social Sciences 07–136, SAGE Publications, Thousand Oaks–London–New Delhi.

Audigier V., Husson F., Josse J. (2016), *Multiple imputation for continuous variables using a Bayesian principal component analysis*, "Journal of Statistical Computation and Simulation", vol. 86, no. 1, pp. 2140–2156, DOI: 10.1080/00949655.2015.1104683.

Blake C., Keogh E., Merz C.J. (1988), *UCI Repository of Machine Learning Datasets*, Department of Information and Computer Science, University of California, Irvine.

Breiman L. (2001), *Random Forests*, "Machine Learning", vol. 45, no. 1, pp. 5–32.

Buuren S. van (2007), *Multiple imputation of discrete and continuous data by fully conditional specification*, "Statistical Methods in Medical Research", vol. 16, no. 3, pp. 219–242.

Buuren S. van (2012), *Flexible Imputation of Missing Data*, Chapman & Hall/CRC Press, Boca Raton–London–New York.

Buuren S. van, Groothuis-Oudshoorn K. (2011), *MICE: Multivariate Imputation by Chained Equations in R*, "Journal of Statistical Software", vol. 45, no. 3, pp. 1–67.

Enders C.K. (2010), *Applied Missing Data Analysis*, The Guilford Press, New York–London.

Hotelling H. (1933), *Analysis of a complex of statistical variables into principal components*, "Journal of Educational Psychology", vol. 24, pp. 417–441, 498–520.

Ilin A., Raiko T. (2010), *Practical Approaches to Principal Component Analysis in the Presence of Missing Values*, "Journal of Machine Learning Research", vol. 11, pp. 1957–2000.

Josse J. (2016), *Contribution to missing values & principal component methods*, Statistics [stat], Université Paris Sud, Orsay.

Josse J., Husson F. (2012), *Handling missing values in exploratory multivariate data analysis methods*, "Journal de la Société Française de Statistique", vol. 153, no. 2, pp. 79–99.

Josse J., Husson F. (2016), *missMDA: A Package for Handling Missing Values in Multivariate Data Analysis*, "Journal of Statistical Software", vol. 70, no. 1, pp. 1–31, DOI: 10.18637/jss.v070.i01.

Josse J., Pagès J., Husson F. (2011), *Multiple imputation in principal component analysis*, "Advances in Data Analysis and Classification", vol. 5, pp. 231–246.

Little R.J.A., Rubin D.B. (2002), *Statistical Analysis with Missing Data*, second edition, Wiley, New Jersey.

Misztal M. (2013), *Some remarks on the data imputation using "missForest" method*, "Acta Universitatis Lodziensis. Folia Oeconomica", vol. 285, pp. 169–179.

Newman D.A. (2014), *Missing Data: Five Practical Guidelines*, "Organizational Research Methods", vol. 17(4), pp. 372–411, DOI: 10.1177/1094428114548590.

Orchard T., Woodbury M.A. (1972), *A missing information principle: Theory and applications*, [in:] *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 697–715.

Pearson K. (1901), *On lines and planes of closest fit to systems of points in space*, "Philosophical Magazine", vol. 6, no. 2, pp. 559–572.

Schafer J.L. (1997), *Analysis of incomplete multivariate data*, Chapman and Hall/CRC, London.

Shah A.D., Bartlett J.W., Carpenter J., Nicholas O., Hemingway H. (2014), *Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study*, "American Journal of Epidemiology", vol. 179, no. 6, pp. 764–774, DOI: 10.1093/aje/kwt312.

Stekhoven D.J., Bühlmann P. (2012), *MissForest – Nonparametric Missing Value Imputation for Mixed-Type Data*, "Bioinformatics", vol. 28, no. 1, pp. 112–118.

Tang F., Ishwaran H. (2017), *Random forest missing data algorithms*, "Statistical Analysis and Data Mining", vol. 10, issue 6, pp. 363–377, DOI: 10.1002/sam.11348.

Wulff J., Ejlskov L. (2017), *Multiple Imputation by Chained Equations in Praxis: Guidelines and Review*, "The Electronic Journal of Business Research Methods", vol. 15, issue 1, pp. 41–56.

Yu L.-M., Burton A., Rivero-Arias O. (2007), *Evaluation of software for multiple imputation of semi-continuous data*, "Statistical Methods in Medical Research", vol. 16, pp. 243–258.

Porównanie wybranych metod imputacji wielokrotnej dla zmiennych ilościowych – wstępne wyniki badań symulacyjnych

Streszczenie: Problem występowania danych niekompletnych i ich wpływu na wyniki analiz statystycznych nie jest związany z żadną konkretną dziedziną nauki – pojawia się w ekonomii, socjologii, edukacji, naukach behawioralnych czy medycynie. W przypadku większości klasycznych metod statystycznych wymagana jest kompletna informacja o zmiennych charakteryzujących badane obiekty, a typowym podejściem do brakujących danych jest po prostu ich usunięcie. Prowadzi to jednak do niewiarygodnych i obciążonych wyników analiz i nie jest zalecane w literaturze przedmiotu. Rekomendowaną metodą postępowania z brakującymi danymi jest imputacja wielokrotna. W artykule rozważono kilka wybranych jej metod. Szczególną uwagę zwrócono na wykorzystanie analizy głównych składowych (PCA) jako metody imputacji. Celem pracy była ocena jakości imputacji opartej na PCA na tle dwóch innych technik uzupełniania braków danych: imputacji wielokrotnej za pomocą równań łańcuchowych (MICE) i metody missForest. Porównania metod imputacji dokonano, wykorzystując podejście symulacyjne i generując braki danych w 10 kompletnych zbiorach danych z repozytorium baz danych Uniwersytetu Kalifornijskiego w Irvine, z uwzględnieniem różnych mechanizmów generowania braków danych oraz różnych proporcji (10–50%) brakujących wartości. Do imputacji brakujących wartości zastosowano metodę równań łańcuchowych, metodę missForest oraz metodę opartą na głównych składowych (MIPCA). Znormalizowany pierwiastek kwadratowy błędu średniokwadratowego (NRMSE) wykorzystano jako miarę dokładności imputacji. Na podstawie przeprowadzonych analiz metoda missForest może być rekomendowana jako ta metoda wielokrotnej imputacji, która zapewnia najwyższą dokładność imputacji braków danych. Imputacja oparta na analizie głównych składowych (PCA) nie prowadzi do zadowalających wyników.

Słowa kluczowe: dane niekompletne, imputacja wielokrotna, analiza głównych składowych, missForest

JEL: C18, C80, C38