A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 285, 2013

Małgorzata Misztal[*]

SOME REMARKS ON THE DATA IMPUTATION USING "MISSFOREST" METHOD

Abstract. Missing data are quite common in practical applications of statistical methods and imputation is a general statistical method for the analysis of incomplete data sets.

Stekhoven and Bühlmann (2012) proposed an iterative imputation method (called "missForest") based on Random Forests (Breiman 2001) to cope with missing values.

In the paper a short description of "missForest" is presented and some selected missing data techniques are compared with "missForest" by artificially simulating different proportions and mechanisms of missing data using complete data sets from the UCI repository of machine learning databases.

Key words: missing values, single and multiple imputation, random forests, missForest.

I. INTRODUCTION

Incomplete data are quite common in practical applications of statistical methods. One way to deal with missing data is to impute all missing values before analysis, using single or multiple imputation methods.

Imputation is the substitution of missing values with some other values in order to obtain the complete data set.

Single imputation consists in filling in missing values once. In multiple imputation – missing values are filled in m times, standard analyses are performed on each of the m imputed data sets and the results from the m analyses are combined into one result.

Another important thing is to understand why the data are missing. According to Little and Rubin (2002) there are three missing data mechanisms: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) and *Not Missing at Random* (NMAR).

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:

[*] Ph.D., Department of Statistical Methods, University of Łódź.

- MCAR means that the probability if an information is missing does not depend on X_{mis} or on X_{obs} ;
- MAR means that the probability if an information is missing does not depend on X_{mis}, but may depend on X_{obs};
- MNAR means that the probability if an information is missing does depend on X_{mis}.

Under an assumption of MCAR or MAR mechanism, to deal with missing data one can use a lot of imputation methods, e. g. mean / mode imputation, conditional mean imputation (regression imputation), stochastic regression imputation, hot deck imputation, substitution, cold deck imputation, maximum likelihood method (ML), EM algorithm, predictive mean matching, k-NN imputation.

NMAR mechanism requires a different and more complex approach, i. e. selection models or pattern-mixture models (see details in Allison 2002 or Little and Rubin 2002).

Another interesting technique for handling missing data is "missForest" – new iterative imputation method proposed by D. J. Stekhoven and P. Bühlmann (2012), which is based on the Breiman's Random Forests (Breiman 2001).

In the paper a short description of "missForest" method is presented and some selected imputation techniques are compared with "missForest" by artificially simulating different proportions and mechanisms of missing data using complete data sets mainly from the UCI repository of machine learning databases.

II. THE IDEA OF MISSFOREST

Let us consider a learning set consisted of n cases characterized by p variables. For an arbitrary variable $\mathbf{X}_s$ with missing values at entries $i_{mis}^{(s)} \in \{1, ..., n\}$ the data set can be divided into 4 parts (see Stekhoven and Bühlmann 2012):

1. The observed values of $\mathbf{X}_s$, denoted by $y_{obs}^{(s)}$;
2. The missing values of $\mathbf{X}_s$, denoted by $y_{mis}^{(s)}$;
3. The variables other than $\mathbf{X}_s$ with observations $i_{obs}^{(s)} \in \{1, ..., n\} \setminus i_{mis}^{(s)}$, denoted by $x_{obs}^{(s)}$;
4. The variables other than $\mathbf{X}_s$ with observations $i_{mis}^{(s)}$, denoted by $x_{mis}^{(s)}$.

Since the index $i_{obs}^{(s)}$ corresponds to the observed values of the variable $\mathbf{X}_s$, $x_{obs}^{(s)}$ can be not completely observed. Also, $x_{mis}^{(s)}$ is typically not completely missing.

Let us analyze some examples. Figure 1 shows an example of data set with missing values.

x1	x2	x3	x4	x5	x6	class
36	702	7	NA	76,34	11	w
32	5000	7	1100	241,49	35	w
31	5000	13	1340,2	292,58	24	w
35	7710	NA	2446	356,53	35	w
32	7679	4	1750	NA	35	w
23	NA	3	560	205,5	18	w
28	1147	NA	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	1320	93	1011,3	142,24	11	w
31	1612	22	594	93,94	NA	w
46	2776	58	NA	272,4	12	s
41	NA	95	1752,2	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	137	2416,2	137,08	NA	s
24	4950	20	1038,9	224,83	35	s
50	1000	34	1163,4	98,13	12	s
42	4742	231	1293	215,39	35	s
NA	3268	105	1940	375,15	10	s
41	3000	214	1539,1	350,55	NA	s
43	3000	71	1433,7	NA	11	s

Figure 1. An example of data set with missing values.
Source: own elaboration.

Figure 2 presents 4 parts of the data set described above, for variable X_6.

x1	x2	x3	x4	x5	x6	class
36	702	7	NA	76,34	11	w
32	5000	7	1100	241,49	35	w
31	5000	13	1340,2	292,58	24	w
35	7710	NA	2446	356,53	35	w
32	7679	4	1750	NA	35	w
23	NA	3	560	205,5	18	w
28	1147	NA	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	1320	93	1011,3	142,24	11	w
31	1612	22	594	93,94	NA	w
46	2776	58	NA	272,4	12	s
41	NA	95	1752,2	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	137	2416,2	137,08	NA	s
24	4950	20	1038,9	224,83	35	s
50	1000	34	1163,4	98,13	12	s
42	4742	231	1293	215,39	35	s
NA	3268	105	1940	375,15	10	s
41	3000	214	1539,1	350,55	NA	s
43	3000	71	1433,7	NA	11	s

Legend:

y_{obs}
y_{mis}
x_{obs}
x_{mis}

Figure 2. $y_{obs}^{(6)}, y_{mis}^{(6)}, x_{obs}^{(6)}, x_{mis}^{(6)}$ for variable X_6.
Source: own elaboration.

According to Stekhoven and Bühlmann (2012), the idea of "missForest" can be described in the following steps:

1. Make initial guess for missing values using mean imputation or any other imputation method.
2. Sort all the variables X_s, s = 1, 2, …, p, according to the amount of missing values, starting with the lowest amount.
3. For each variable X_s fit a Random Forest with response $y_{obs}^{(s)}$ and predictors $s_{obs}^{(s)}$. Then, predict the missing values $y_{mis}^{(s)}$ by applying the trained random forest to $x_{mis}^{(s)}$.
4. The imputation procedure is repeated until a stopping criterion γ is met.

The stopping criterion γ is met as soon as the difference between the newly imputed data matrix and the previous one increases for the first time with respect to both variable types (continuous and categorical), if present.

The difference for the set of continuous variables N is defined as:

$$\Delta_N = \frac{\sum_{j \in N} \left(X_{new}^{imp} - X_{old}^{imp} \right)^2}{\sum_{j \in N} \left(X_{new}^{imp} \right)^2}, \tag{1}$$

where: X_{new}^{imp} and X_{old}^{imp} denote new and previously imputed data matrix, respectively.

The difference for the set of categorical variables F is defined as:

$$\Delta_F = \frac{\sum_{j \in F} \sum_{i=1}^{n} I_{X_{new}^{imp} \neq X_{old}^{imp}}}{\# NA}, \tag{2}$$

#NA is the number of missing values in the categorical variables.

The performance of the method can be assessed using NRMSE (*normalised root mean squared error*) proposed by Oba *et al*. (2003). For continuous variables it is defined as:

$$NRMSE = \sqrt{\frac{mean((X^{true} - X^{imp})^2)}{\text{var}(X^{true})}}, \tag{3}$$

where:

X^{true} – complete data set;

X^{imp} – imputed data set;

mean, *var* – empirical mean and variance computed over the continuous missing values.

Stekhoven and Bühlmann (2012) compared the "missForest" method to kNN imputation (Troyanskaya *et al*. (2001)), MissPALasso (a method based on EM algorithm, proposed by Städler and Bühlmann (2010)) and MICE (van Buuren S and Groothuis-Oudshoorn (2011)). They showed that "missForest" could outperform other imputation methods. Let us observe, however, that in simulation experiments only the missing completely at random data were analyzed. It is reasonable, therefore, to carry out additional experiments to assess the usefulness of the "missForest" imputation method.

III. SIMULATION EXPERIMENTS

In order to compare the "missForest" method with other imputation techniques 11 complete data sets from the UCI repository of machine learning databases (Blake *et al*. 1988) and from author's research (AR) were selected. Short description of all the data sets is presented in Tab. 1.

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

Data set	Id	Source	Number of cases	Number of predictors (all continuous)	Number of classes
Protein Localization Sites	E.coli	UCI	336	5	8
Glass Identification Database	glass	UCI	214	9	2
Haberman's Survival Data	haberman	UCI	306	3	2
Iris Plants Database	iris	UCI	150	4	3
BreastTissue	breastT	UCI	106	9	6
Wine recognition data	wine	UCI	178	13	3
Wisconsin Prognostic Breast Cancer	wpbc	UCI	194	12	2
Vertebral Column	vertebral	UCI	310	6	2
Borrowers	cred	AR	100	6	2
Drug Addicts	drug	AR	60	5	2
Metabolic Syndrome	ms	AR	86	21	2

Source: own elaboration.

Missing data were applied into each data set assuming the general missing data pattern and 3 mechanisms of missing data – MCAR, MAR, NMAR.

Under the MCAR assumption missing values were randomly applied into each data set.

Under the MAR assumption, probability of information being missing depended on class attribute.

Under the NMAR assumption, the biggest or the smallest values of X_s were removed.

An example of complete data set and the results of introducing missing data according to different missing data mechanisms is given in Figure 3.

Five levels of proportion of missingness were considered: 5%, 10%, 20%, 30%, 40%. The following imputation methods were taken into account:

– Mean imputation (*mean*);
– Hot deck imputation – missing values were imputed using sampling with replacement from the observed data (*sample*);
– Predictive mean matching (*pmm*);
– "missForest" (*mF*).

Complete data set:

x1	x2	x3	x4	x5	x6	class
36	702	7	850	76,34	11	w
32	5000	7	1100	241,49	35	w
31	5000	13	1340,2	292,58	24	w
35	7710	24	2446	356,53	35	w
32	7679	4	1750	370,88	35	w
23	2800	3	560	205,5	18	w
28	1147	4	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	1320	93	1011,3	142,24	11	w
31	1612	22	594	93,94	23	w
46	2776	58	1034,3	272,4	12	s
41	2199	95	1752,2	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	137	2416,2	137,08	35	s
24	4950	20	1038,9	224,83	35	s
50	1000	34	1163,4	98,13	12	s
42	4742	231	1293	215,39	35	s
32	3268	105	1940	375,15	10	s
41	3000	214	1539,1	350,55	10	s
43	3000	71	1433,7	323,28	11	s

MCAR:

x1	x2	x3	x4	x5	x6	class
36	702	7	NA	76,34	11	w
32	5000	7	1100	241,49	35	w
31	5000	13	1340	292,58	24	w
35	7710	NA	2446	356,53	35	w
32	7679	4	1750	NA	35	w
23	NA	3	560	205,5	18	w
28	1147	NA	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	1320	93	1011	142,24	11	w
31	1612	22	594	93,94	NA	w
46	2776	58	NA	272,4	12	s
41	NA	95	1752	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	137	2416	137,08	NA	s
24	4950	20	1039	224,83	35	s
50	1000	34	1163	98,13	12	s
42	4742	231	1293	215,39	35	s
NA	3268	105	1940	375,15	10	s
41	3000	214	1539	350,55	NA	s
43	3000	71	1434	NA	11	s

MAR:

x1	x2	x3	x4	x5	x6	class
36	702	7	850	76,34	11	w
32	NA	7	1100	241,49	35	w
31	NA	13	1340,2	NA	24	w
35	7710	24	2446	356,53	35	w
NA	7679	4	1750	370,88	35	w
23	2800	3	NA	205,5	18	w
28	1147	4	676,9	114,91	12	w
20	1300	5	NA	NA	23	w
36	1320	93	1011,3	142,24	11	w
31	1612	22	594	93,94	NA	w
46	2776	58	1034,3	272,4	12	s
NA	2199	95	1752,2	154,7	18	s
23	4300	8	1450	195,31	35	s
40	3018	NA	2416,2	137,08	35	s
NA	NA	20	1038,9	224,83	35	s
50	1000	34	1163,4	98,13	12	s
42	4742	231	1293	215,39	35	s
32	3268	105	1940	375,15	10	s
41	3000	214	1539,1	350,55	10	s
43	3000	71	1433,7	323,28	11	s

NMAR:

x1	x2	x3	x4	x5	x6	class
36	702	7	850	76,34	11	w
32	NA	7	1100	241,49	35	w
31	NA	13	1340	292,58	24	w
35	NA	24	2446	356,53	35	w
32	NA	4	1750	370,88	35	w
23	NA	3	560	205,5	18	w
28	1147	4	676,9	114,91	12	w
20	1300	5	1950	77,36	23	w
36	NA	93	1011	142,24	11	w
31	NA	22	594	93,94	23	w
46	2776	58	1034	272,4	12	s
NA	2199	95	1752	154,7	18	s
NA	4300	8	1450	195,31	35	s
NA	3018	137	2416	137,08	35	s
NA	4950	20	1039	224,83	35	s
50	1000	34	1163	98,13	12	s
42	4742	231	1293	215,39	35	s
NA	3268	105	1940	375,15	10	s
41	3000	214	1539	350,55	10	s
43	3000	71	1434	323,28	11	s

Figure 3. An example of complete data set and data sets with MCAR, MAR and NMAR missing values

Source: own elaboration.

Since all the predictors in the analyzed data sets were continuous, NRMSE was calculated to assess the quality of imputation.

All the calculations were performed using the R environment with two packages: `missForest` and `mice`.

For 11 data sets, 3 missing data mechanisms, 5 levels of proportion of missingness and 4 imputation methods, the final NRMSE is averaged over the 1000 repetitions.

IV. RESULTS AND CONCLUDING REMARKS

The results are summarized using the box-and-whiskers plots (median/IQR/min-max, including outliers) in Figures 4–6.

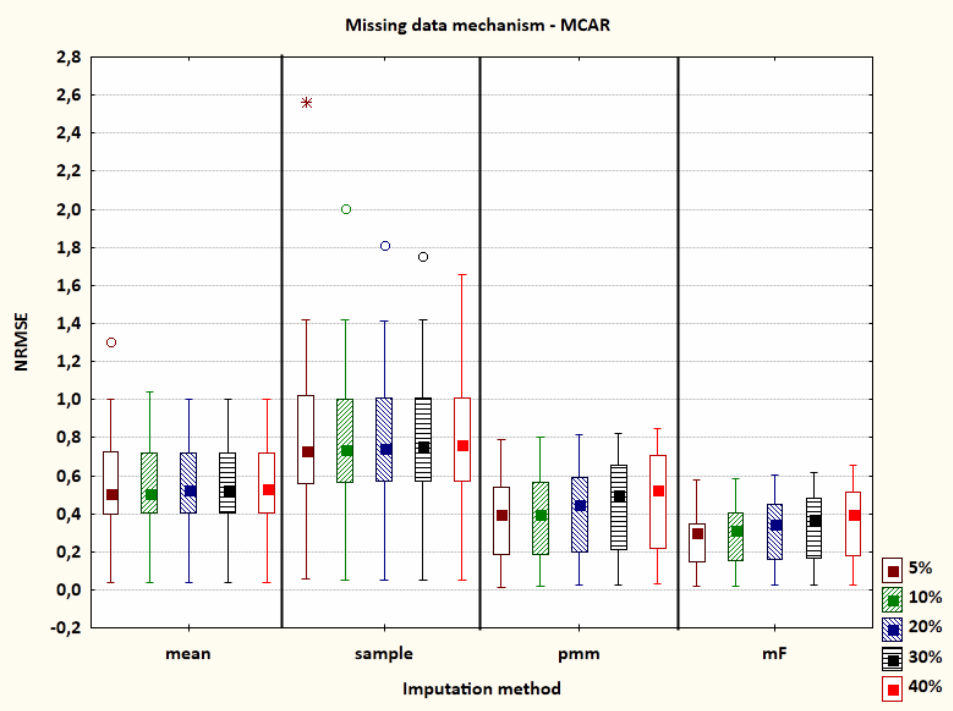

Figure 4. Comparison of the results for MCAR data

Source: own calculations.

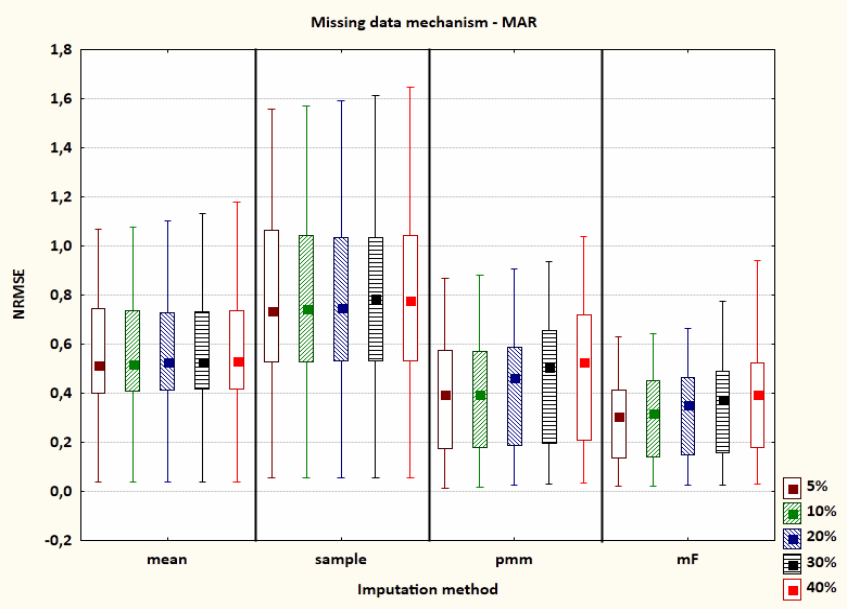

Figure 5. Comparison of the results for MAR data

Source: own calculations.

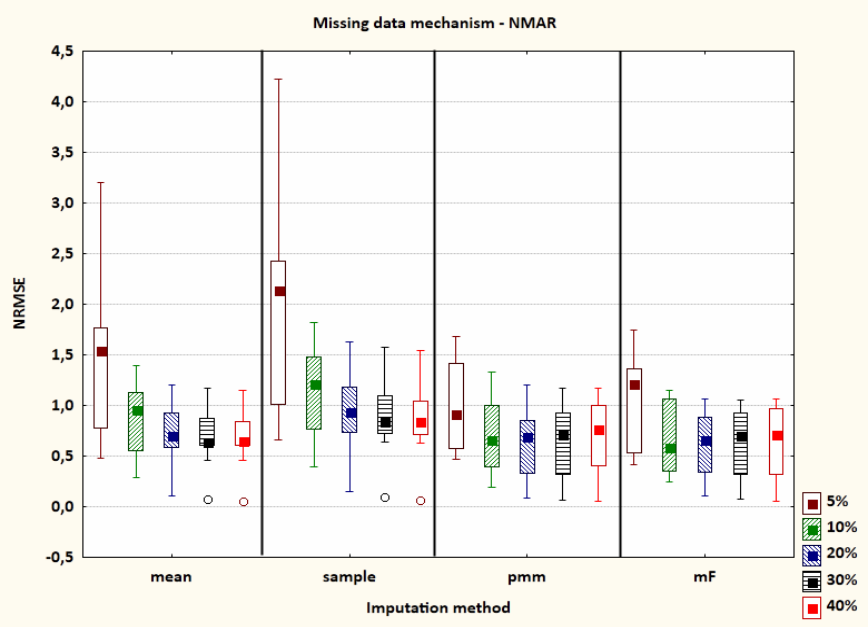

Figure 6. Comparison of the results for NMAR data

Source: own calculations.

As one can see, there is a correlation between the results (NRMSE) and the missing data mechanism. NRMSE is smaller for MCAR and MAR data compared with NMAR data.

The "missForest" imputation outperforms all the other methods in the case of randomly missing data (MCAR or MAR). In the case of NMAR all the errors are much bigger and the advantage of "missForest" is not so spectacular.

The comparison of the selected imputation methods is also showed in Figures 7–9, where the decrease of NRMSE (in %) for "missForest" method is presented.

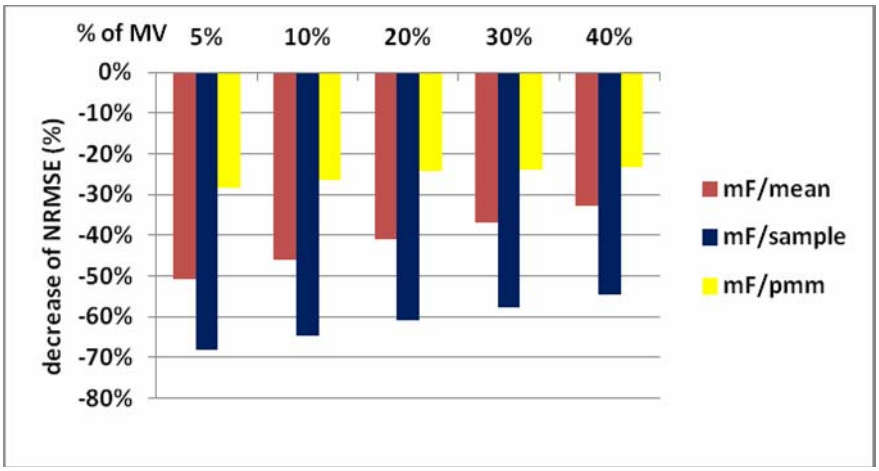

Figure 7. Decrease of NRMSE for MCAR data

Source: own calculations.

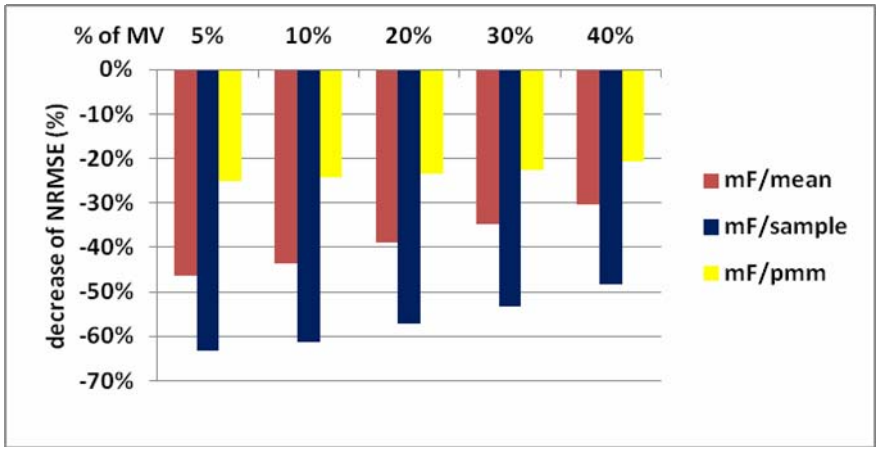

Figure 8. Decrease of NRMSE for MAR data

Source: own calculations.

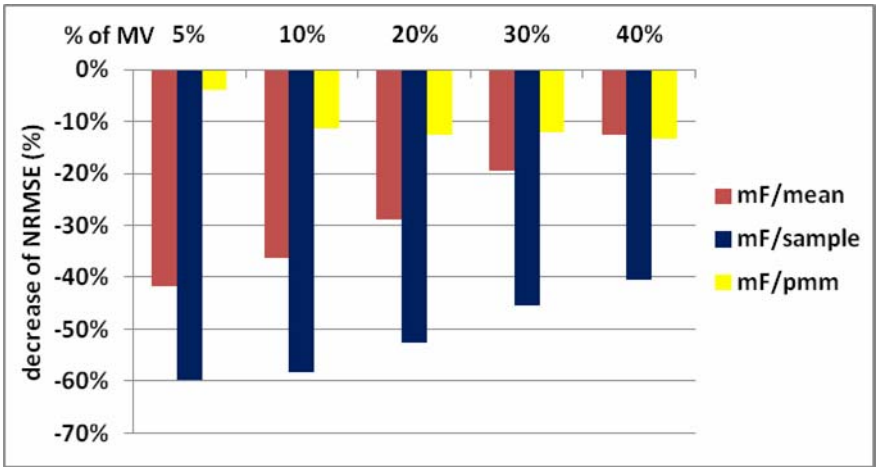

Figure 9. Decrease of NRMSE for NMAR data

Source: own calculations.

The decrease of NRMSE is the highest for "missForest" compared to sample imputation and the smallest for "missForest" compared to predictive mean matching.

The bigger the percentage of missing values the smaller the decrease of NRMSE. Differences between all the analyzed imputation methods are less evident for NMAR data.

On the other hand, NRMSE seems to be inappropriate to assess the quality of imputation, especially for NMAR missing data mechanism. If the variance is small, the error increases; such situation is especially frequent for NMAR data and small percentage of missing values (see Fig. 6).

All the results presented should be viewed as an initial step to more complex analysis of the "missForest" method. Some other imputation methods and measures will be proposed and tested in further research.

REFERENCES

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

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" 45(1): 5–32.

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

Oba S., Sato M., Takemasa I., Monden M., Matsubara K., Ishii S. (2003), *A Bayesian Missing Value Estimation Method for Gene Expression Profile Data*, "Bioinformatics" 19(16): 2088–2096.

Städler N., Bühlmann P. (2010), *Pattern Alternating Maximization Algorithm for High-Dimensional Missing Data*, Arxiv preprint arXiv:1005.0366.

Stekhoven D. J., Bühlmann P. (2012), *MissForest – Nonparametric Missing Value Imputation for Mixed-Type Data*, "Bioinformatics" 28(1): 112–118.

Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R. (2001), *Missing Value Estimation Methods for DNA Microarrays*, "Bioinformatics" 17(6): 520–525.

van Buuren S., Groothuis-Oudshoorn K. (2011), *MICE: Multivariate Imputation by Chained Equations in R*, „Journal of Statistical Software", 45(3): 1–67.

Małgorzata Misztal

KILKA UWAG O IMPUTACJI DANYCH Z WYKORZYSTANIEM METODY "MISSFOREST"

W pracy Stekhovena i Bühlmanna (2012) zaproponowano nową iteracyjną metodę imputacji (nazwaną „missForest") opartą na metodzie *Random Forests* Breimana (2001).

W niniejszym artykule omówiono metodę „missForest" i porównano kilka wybranych technik postępowania w sytuacji występowania braków danych z metodą „missForest". W tym celu wykorzystano podejście symulacyjne generując różne proporcje i mechanizmy powstawania braków danych w zbiorach danych pochodzących głównie z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine.