Multiple imputation through machine learning in a survey of sport clubs

Organisation:	Statistics Poland
Author(s):	Sebastian Wójcik, Agnieszka Giemza
Date:	04-10-2021
Version:	1.0

1. Background [a short description of problem you want to solve with ML and reasons why you want to use ML]

Statistics Poland conducts a census survey of activity of sport clubs which takes place every two years. Survey frame contains 16432 sport clubs of which 1660 (10,1%) requires imputation The data is disseminated on the highest level of spatial disaggregation. Hence the data processing must take that into account.
Methodological Committee (advisory body of president of Statistics Poland) obliged our office to implement multiple imputation in this survey. Thus, we shall test several ML and non-ML methods at multiple imputation task under this project to meet the obligation.

2. Data

2.1 Input data[a short description of input data, an example of how typical data record looks like would be helpful]

Data contains 48 variables of which 44 are count data and 4 are categorical data. The first part of the database pertains to general information about sport club such as members of sport club, persons practising sports, competitors, coaching staff - with respect to age and gender. It can be tackled quite easy with standard methods. The second part of the database covers sport disciplines. Number of them varies among sport clubs and if missing need to be imputed. That part is not covered with this report.

2.2 Data preparation[if there was any data preparation (e.g. data cleaning, text normalisation)]

Not needed. Additional variables e.g. NUTS1-NUTS3 code, locality code and sports association are taken from the survey frame to enhance model results.

2.3 Feature selection[if there was any feature selection]

At start, crucial variables were selected:

- Members of sports club
- Persons practising sports
- Competitors

Such a small subset of variables allowed to test the code quickly without hardware burden. Next, other variables from the survey were added to the database. Finally, the database contained 7,000 observations and 13 variables. The table below presents a short description of the variables and their codes in the input file.

Tab. 1. Variables

CODE	VARIABLE DESCRIPTION	TYPE
D1W2	Indicates which sport association does the sport club belong to	categorical
D1W3	Indicates if the sport club participates in the sports competitions	categorical
D2W1	Members of sports club	numeric
D2W2	Persons practising sports	numeric
D2W3	Men practising sports	numeric
D2W4	Men practising sports under 18 years old	numeric
D2W5	Women practising sport	numeric
D2W6	Woman practising sport under 18 years old	numeric
D2W7	Competitors registered in Polish or district sport association	numeric
D2W8	Male competitors	numeric
D2W9	Male competitors under 18 years old	numeric
D2W10	Female competitors	numeric
D2W11	Female competitors under 18 years old	numeric

2.4 Output data [a short description of how output data looks like]

For each imputation method and set of parameters, we ran 300 simulations and calculated precision measures such as MAE, RMSE, Accuracy and R^2. Output data contains precision measures for all simulations.

3. Machine learning solution

3.1 Models

We tested the following methods:

- MissForest- Nonparametric Missing Value Imputation using Random Forest
- MICE with CART (Classification and Regression Trees)
- MICE with PMM (Predictive Mean Matching)
- MICE with BLR (Bayesian Linear Regression).

MissForest is not a "pure" multiply imputation method as proposed by D. Rubin in 1987. It does not deliver several imputation sets. Nevertheless, it is advocated that it can handle an imputation of mixed data (numeric and factor variables in one data frame). It has an advantage over some other methods that cannot deal both with regression and classification problem. Multivariate Imputation By Chained Equations (MICE) is considered as a principled method of dealing with missing data. The first step of MICE requires a single imputation method to fill in the missing data, the first guess. Since the selection of a single imputation method is arbitrary, we selected three methods: Bayesian Linear Regression (definitely non-ML method), Classification and Regression Trees (definitely ML method) and Predictive Mean Matching (rather non-ML method or hybrid).

3.2 Model(s) finally selected and quality criteria used (e.g. accuracy, time)[which model was selected? What quality measures were used to compare different ML models (e.g. accuracy (e.g. RMSE, MAE, F1, precision), runtime to train the model (e.g.2 hours for 500,000 training samples and 25 features))]

Let us start with some notation. Assume that $\mathrm{X} = (x_1, \dots, x_n)$ is the vector of true values and let $\mathrm{X}^* = (x_1^*, \dots, x_n^*)$ is the vector of imputed values. For categorical variables (D1W2, D1W3) we calculated accuracy

$$ACC(X) = \sum_{I=1}^{n} Ind(x),$$

where

$$Ind(x) = \begin{cases} 1 & if \quad x_i = x_i^* \ \ i = 1, \dots, n \\ 0 & if \quad x_i \neq x_i^* \ i = 1, \dots, n \end{cases}.$$

For numerical variables D2W1-D1W11 we calculated:

- Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |x_i - x_i^*|,$$

- Root of Mean Square Error

$$RMSE = \sqrt{\sum_{i=1}^{n} (x_i - x_i^*)^2},$$

- Coefficient of determination (squared Pearson's correlation coefficient).

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})\,(x_i^* - \bar{x}^*)\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}\left(x_i^* - \bar{x}_i^*\right)^2}.$$

It should be noticed this definition works well when imputation is unbiased, that is $\bar{x} = \bar{x}^*$.

We checked two conditions, that is if D2W2=D2W3+D2W5 and D2W7=D2W8+D2W10. The left-hand side of each equality was treated as true value while the right-hand side of each equality was treated as imputed value. It allowed to calculate aforementioned precision measures.

We also checked if the following inequalities hold:

- men practising sports $\geq$ men practising sports under 18 years old (D2W3 $\geq$D2W4)
- women practising sports $\geq$ women practising sports under 18 years old (D2W5 $\geq$D2W6)
- male competitors $\geq$ male competitors under 18 years old (D2W8 $\geq$D2W9)
- female competitors $\geq$ female competitors under 18 years old (D2W10 $\geq$D2W11)

For each of them, we calculated the percentage of imputed data such that the given equality holds.

The second criterion of comparison was runtime. The next table shows runtime of simulations for each method. Details of simulation procedure are presented in the next section.

Tab. 2. Runtime of 300 simulations (7000 training samples and 13 features) with respect to the share of missing data

Method	Share of missing data				
	10%	20%	30%	40%	50%
MissForest	13h28min	13h 29min	13h30min	13h30min	13h31min
MICE with PMM	47 min	46 min	39 min	40 min	40 min
MICE with CART	8h 12 min	8h 15 min	8h 20min	8h 17min	8h 8min

MICE with Bayesian Linear Regression	34 min	35 min	36 min	38 min	42 min

Methods were compared with respect to precision and stability of results, distributional properties and runtime under two different assumption on mechanism of generating missing data. In a case when missing data pattern was Missing Completely at Random (MCAR) all methods achieved quite similar results. However, when missing data pattern was Missing Not at Random (MNAR), precision of the results varied. Moreover, MICE with Bayesian Linear Regression produced unstable and inadmissible results. In terms of precision, MissForest and MICE with CART achieved the best results. Both methods preserved four examined inequalities more often than other two methods. Also, ML methods achieved better distributional properties. The only disadvantage of ML methods is runtime.

Taking into account all arguments, MissForest was chosen for further implementation in our survey.

3.3Hardware used[e.g. Intel Core i5-6300U, 2.4GHz]

Intel Core i5-9400 CPU, 2.90GHz

4. Results[result of applying the selected model; if possible, please provide quantitative measures comparing with existing / status-quomethods in terms of accuracy (e.g. manual coding had 0.80 precision), time (e.g. 4,000 hours for manual coding), cost, etc.]

Independent of size of imputed data.

In this study we examined two assumptions on mechanism of generating missing data. Under Missing Completely at Random (MCAR) assumption, the probability of being missing is the same only within groups defined by the observed data. Missing Not at Random (MNAR) means that the probability of being missing varies for reasons that are unknown to us and may depend on unobservable data. In a case of MCAR, a given percentage of data was removed at random. In the latter case, all variables were removed except totals, that is, block of variables were removed while crucial variables were left. It is the most common case, when interviewee provides total value without breakdown. We tested a share of 10, 20, 30, 40 and 50% of missing data to follow how precision measures changed with decreasing training set. For each mechanism of generating missing data and each share of missing data, 300 simulations were carried out. Some variables from statistical frame were always available.

The following tables show the mean results of MAE and R^2 for each method. Accuracy was calculated for the variables of the factor type.

Tables 3-6 show the results under MCAR assumption.

Tab. 3. MISSFOREST - MCAR

	MAE					**R^2 / ACC***				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D1W2*	-	-	-	-	-	0,5971	0,5918	0,5850	0,5750	0,5597
D1W3*	-	-	-	-	-	0,7160	0,7117	0,7059	0,6983	0,6878
D2W1	21,045	20,884	21,143	21,945	23,484	0,6465	0,6376	0,6230	0,6087	0,5766
D2W2	4,858	5,943	7,392	9,317	11,965	0,9450	0,9198	0,8910	0,8562	0,8144
D2W3	4,663	5,702	6,945	8,601	10,780	0,9368	0,9033	0,8623	0,8212	0,7791
D2W4	7,635	8,596	9,657	10,965	12,642	0,8891	0,8606	0,8308	0,7911	0,7434
D2W5	3,325	3,954	4,721	5,762	7,097	0,8960	0,8469	0,8031	0,7457	0,6769

D2W6	3,422	3,955	4,585	5,444	6,542	0,8652	0,8216	0,7702	0,7086	0,6331
D2W7	3,772	4,895	6,318	8,190	10,713	0,9471	0,9046	0,8574	0,7960	0,7452
D2W8	3,491	4,305	5,540	7,015	9,238	0,9323	0,8923	0,8404	0,7847	0,7194
D2W9	3,936	4,747	5,618	6,759	8,322	0,9404	0,9117	0,8837	0,8412	0,7797
D2W10	1,484	1,791	2,218	2,766	3,561	0,9495	0,9237	0,8856	0,8308	0,7515
D2W11	1,262	1,503	1,862	2,320	2,983	0,9287	0,9040	0,8610	0,8021	0,7127

Tab. 4. MICE with Bayesian Linear Regression - MCAR

	MAE					**R^2 / ACC***				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D1W2*	-	-	-	-	-	0,4107	0,4091	0,4083	0,4059	0,4042
D1W3*	-	-	-	-	-	0,5263	0,5251	0,5219	0,5192	0,5157
D2W1	36,335	36,745	37,315	36,745	36,335	0,6350	0,6031	0,5875	0,5604	0,5344
D2W2	4,153	7,225	10,811	7,225	4,153	0,9724	0,9541	0,9241	0,8892	0,8388
D2W3	3,370	6,070	9,216	6,070	3,370	0,9651	0,9430	0,9088	0,8713	0,8195
D2W4	11,602	13,664	15,691	13,664	11,602	0,8754	0,8330	0,7876	0,7186	0,6413
D2W5	2,523	3,931	5,770	3,931	2,523	0,9244	0,9012	0,8612	0,8014	0,7131
D2W6	6,098	6,857	7,597	6,857	6,098	0,8822	0,8428	0,7962	0,7329	0,6550
D2W7	3,100	5,521	8,693	5,521	3,100	0,9708	0,9499	0,9185	0,8703	0,8166
D2W8	2,711	5,062	7,967	5,062	2,711	0,9645	0,9400	0,9081	0,8537	0,7947
D2W9	10,228	11,711	13,312	11,711	10,228	0,8560	0,8142	0,7625	0,6962	0,6211
D2W10	1,280	1,972	3,180	1,972	1,280	0,9455	0,9236	0,8818	0,8054	0,6944
D2W11	3,043	3,438	3,943	3,438	3,043	0,9030	0,8720	0,8328	0,7629	0,6621

Tab. 5. MICE with CART - MCAR

	MAE					**R^2 / ACC***				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D1W2*	-	-	-	-	-	0,5706	0,5625	0,5850	0,5750	0,5597
D1W3*	-	-	-	-	-	0,6971	0,6941	0,7059	0,6983	0,6878
D2W1	24,620	25,140	21,140	21,940	23,480	0,6039	0,5608	0,6230	0,6087	0,5766
D2W2	8,085	9,971	7,392	9,317	11,960	0,9151	0,8737	0,8910	0,8562	0,8144
D2W3	7,941	9,268	6,945	8,601	10,780	0,9056	0,8650	0,8623	0,8212	0,7791
D2W4	9,235	10,400	9,657	10,960	12,640	0,8832	0,8443	0,8308	0,7911	0,7434
D2W5	5,039	5,944	4,721	5,762	7,097	0,8265	0,7727	0,8031	0,7457	0,6769
D2W6	4,290	5,149	4,585	5,444	6,542	0,8391	0,7803	0,7702	0,7086	0,6331
D2W7	5,929	7,337	6,318	8,190	10,710	0,9112	0,8590	0,8574	0,7960	0,7452
D2W8	5,549	6,840	5,540	7,015	9,238	0,8988	0,8278	0,8404	0,7847	0,7194
D2W9	4,601	5,572	5,618	6,759	8,322	0,9260	0,8858	0,8837	0,8412	0,7797
D2W10	2,084	2,414	2,218	2,766	3,561	0,8918	0,8709	0,8856	0,8308	0,7515
D2W11	1,487	1,819	1,862	2,320	2,983	0,8885	0,8500	0,8610	0,8021	0,7127

Tab. 6. MICE with Bayesian Linear Regression - MCAR

	MAE					**R^2 / ACC***				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D1W2*	-	-	-	-	-	0,4313	0,4272	0,4236	0,4201	0,4162
D1W3*	-	-	-	-	-	0,6261	0,6172	0,6097	0,5988	0,5872
D2W1	71,823	67,857	71,095	70,763	72,718	0,4059	0,3837	0,3681	0,3511	0,3248
D2W2	18,626	27,746	38,888	43,927	53,104	0,8685	0,7921	0,7048	0,6601	0,5715

D2W3	18,027	25,849	31,770	34,777	41,086	0,7946	0,7011	0,6385	0,6070	0,5318
D2W4	30,537	29,503	31,139	32,576	39,974	0,6226	0,6232	0,5963	0,5523	0,4680
D2W5	12,141	14,945	18,858	21,619	24,740	0,8138	0,7436	0,6516	0,5730	0,4848
D2W6	14,495	14,718	17,259	18,898	21,378	0,7303	0,6852	0,6240	0,5523	0,4484
D2W7	14,604	23,088	30,620	34,956	42,528	0,8552	0,7599	0,6752	0,6026	0,5261
D2W8	14,267	21,617	25,032	32,168	34,085	0,8074	0,7079	0,6526	0,5548	0,5077
D2W9	24,107	24,480	27,556	28,153	29,177	0,6438	0,6276	0,5707	0,5453	0,4925
D2W10	6,586	8,348	9,299	11,398	12,275	0,8644	0,7838	0,7110	0,5974	0,5044
D2W11	7,660	8,010	9,632	9,411	11,524	0,7799	0,7239	0,6427	0,5581	0,4504

All methods achieved quite similar results with a little advantage of ML methods.

Tables 7-10 show the results under MNAR assumption.

Tab. 7. MISSFOREST - MNAR

	MAE					**R^2**				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D2W3	16,143	16,167	16,160	16,129	16,141	0,7274	0,7252	0,7261	0,7273	0,7252
D2W4	17,661	17,660	17,670	17,653	17,659	0,6577	0,6572	0,6569	0,6576	0,6573
D2W5	15,371	15,368	15,356	15,343	15,351	0,4415	0,4414	0,4420	0,4414	0,4416
D2W6	14,321	14,326	14,317	14,301	14,306	0,3301	0,3293	0,3287	0,3284	0,3288
D2W8	11,560	11,574	11,534	11,516	11,538	0,7267	0,7253	0,7254	0,7271	0,7252
D2W9	12,394	12,402	12,417	12,404	12,405	0,7006	0,7003	0,7007	0,7006	0,7011
D2W10	9,987	9,994	9,989	9,972	9,970	0,3058	0,3042	0,3044	0,3040	0,3032
D2W11	8,862	8,873	8,865	8,852	8,851	0,2182	0,2153	0,2161	0,2157	0,2149

Tab. 8. MICE with CART - MNAR

	MAE					**R^2**				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D2W3	21,395	21,406	21,439	21,421	21,367	0,4918	0,4805	0,4853	0,4815	0,4845
D2W4	20,815	20,678	20,753	20,670	20,668	0,4888	0,4654	0,4748	0,4652	0,4672
D2W5	21,693	21,633	21,687	21,653	21,660	0,0465	0,0428	0,0446	0,0420	0,0423
D2W6	18,325	18,186	18,266	18,195	18,200	0,0361	0,0341	0,0351	0,0336	0,0340
D2W8	13,158	13,263	13,299	13,321	13,228	0,5068	0,4960	0,4982	0,4994	0,5002
D2W9	13,208	13,338	13,264	13,338	13,329	0,5645	0,5457	0,5555	0,5446	0,5475
D2W10	10,759	10,761	10,795	10,762	10,758	0,0533	0,0493	0,0502	0,0487	0,0494
D2W11	8,615	8,582	8,628	8,582	8,579	0,0427	0,0389	0,0392	0,0380	0,0388

Tab. 9. MICE with Bayesian Linear Regression - MNAR

	MAE					**R^2**				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D2W3	21,744	22,452	21,797	21,928	22,018	0,8355	0,8457	0,8318	0,8168	0,8146
D2W4	1E+11	1E+09	30,232	55,017	746,202	0,0101	0,2593	0,4058	0,1769	0,0113
D2W5	21,744	22,452	21,797	21,928	22,018	0,0002	0,0007	0,0002	0,0002	0,0003
D2W6	2E+12	2E+08	20,136	23,544	35,760	0,0076	0,0252	0,0169	0,0967	0,0020
D2W8	12,069	12,384	12,114	12,200	12,334	0,9055	0,9123	0,9097	0,8992	0,9014
D2W9	6E+12	7E+08	24,896	43,290	552,930	0,0065	0,2790	0,4196	0,1761	0,0139
D2W10	12,069	12,384	12,114	12,200	12,334	0,0002	0,0007	0,0002	0,0002	0,0003

D2W11	6E+13	1E+10	11,399	13,091	30,907	0,0039	0,0478	0,0336	0,0951	0,0019

Tab. 10. MICE with PMM - MNAR

	MAE					**R^2**				
	10%	**20%**	**30%**	**40%**	**50%**	**10%**	**20%**	**30%**	**40%**	**50%**
D2W3	17,780	17,931	17,782	17,781	17,786	0,7411	0,7401	0,7417	0,7417	0,7412
D2W4	109,747	104,900	111,004	106,054	101,921	0,0737	0,1144	0,0723	0,0933	0,0908
D2W5	17,409	17,424	17,413	17,413	17,419	0,3424	0,3446	0,3430	0,3418	0,3415
D2W6	21,950	23,390	22,013	21,859	21,869	0,1025	0,1075	0,1027	0,1061	0,1063
D2W8	9,298	9,299	9,292	9,286	9,295	0,7493	0,7497	0,7496	0,7503	0,7497
D2W9	98,188	91,819	101,429	93,816	94,318	0,0573	0,1059	0,0524	0,0785	0,0715
D2W10	8,791	8,788	8,784	8,781	8,787	0,1798	0,1798	0,1800	0,1793	0,1793
D2W11	10,696	11,420	10,723	10,706	10,530	0,0399	0,0469	0,0399	0,0422	0,0421

In a case when missing data pattern was Missing Not at Random (MNAR), precision of the results varied. Moreover, MICE with Bayesian Linear Regression produced unstable and inadmissible results. Imputed data often was completely out of range of real data. MICE with PMM also produced high errors in a case of several variables. Such anomalies did not occur in outputs of MissForest and MICE with CART. In terms of precision, in the most of cases, MissForest and MICE with CART achieved the best results with respect to each share of missing data for each variable.

We checked if two equalities: D2W2 = D2W3 + D2W5 (Practitioners) and D2W7 = D2W8 + D2W10 (Competitors), held after imputation. The next table presents the results when 30% of data is missing.

Tab. 11. Results on preserving equalities.

MCAR		**MISSFOREST**	**MICE PMM**	**MICE CART**	**MICE BRL**
MAE	Practitioners	4,716	35,624	4,716	1,196
	Competitors	3,292	26,611	3,292	0,763
RMSE	Practitioners	28,963	80,508	28,963	1,916
	Competitors	24,368	62,864	24,368	1,252
R^2	Practitioners	0,9199	0,6757	0,9199	0,9996
	Competitors	0,9034	0,6666	0,9034	0,9997
MNAR		**MISSFOREST**	**MICE PMM**	**MICE CART**	**MICE BRL***
MAE	Practitioners	5,709	0,782	16,586	0
	Competitors	4,903	0,563	9,502	0
RMSE	Practitioners	42,274	28,404	105,600	0
	Competitors	34,405	28,997	80,259	0
R^2	Practitioners	0,8664	0,9356	0,5399	1
	Competitors	0,8384	0,8830	0,5745	1

*MAE and RMSE rounded to three digits, R^2 rounded to four digits

Non-ML methods obtained better results in a case of MNAR with outstanding performance of MICE with BLR. MICE with PMM achieved poor results under MCAR. Nevertheless, preserving equalities is not a crucial criterion since such equalities can be introduced after the multiply imputation is performed.

The next two tables show the percentage of imputed records such that the following inequalities held

- men practising sports ≥ men practising sports under 18 years old (D2W3 ≥D2W4)
- women practising sports ≥ women practising sports under 18 years old (D2W5 ≥D2W6)
- male competitors ≥ male competitors under 18 years old (D2W8 ≥D2W9)
- female competitors ≥ female competitors under 18 years old (D2W10 ≥D2W11)

Tab. 12. Results on preserving inequalities under MCAR.

	MISSFOREST	MICE PMM	MICE CART	MICE BRL
men practising sports	0,9704	0,7678	0,9704	0,8085
women practising sports	0,9519	0,6793	0,9519	0,6956
male competitors	0,9648	0,7032	0,9648	0,7398
female competitors	0,9681	0,6922	0,9681	0,7246

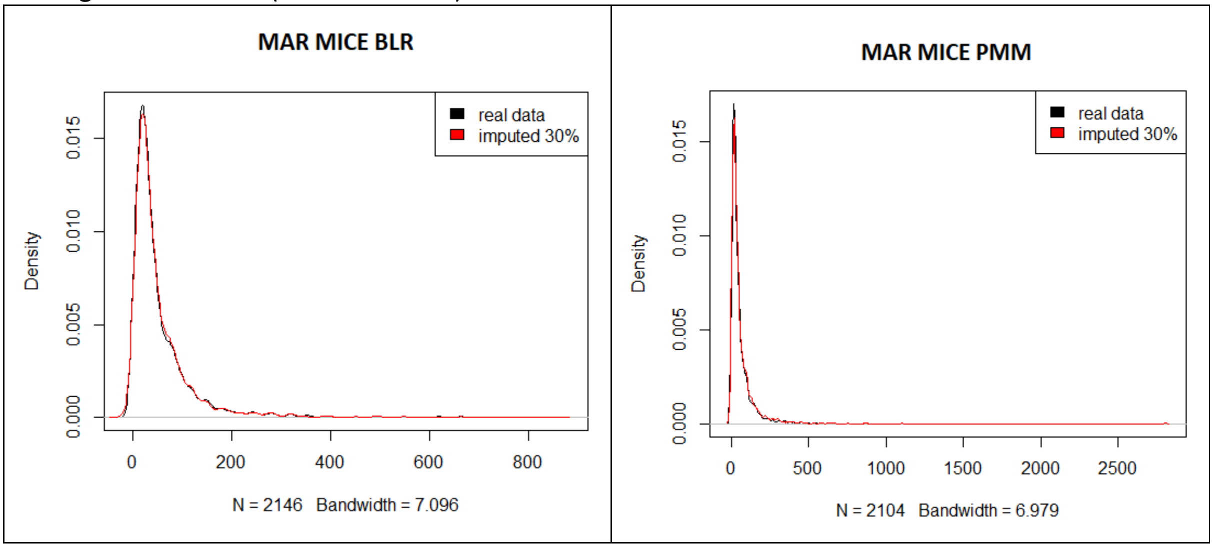
Tab. 13. Results on preserving inequalities under MNAR.

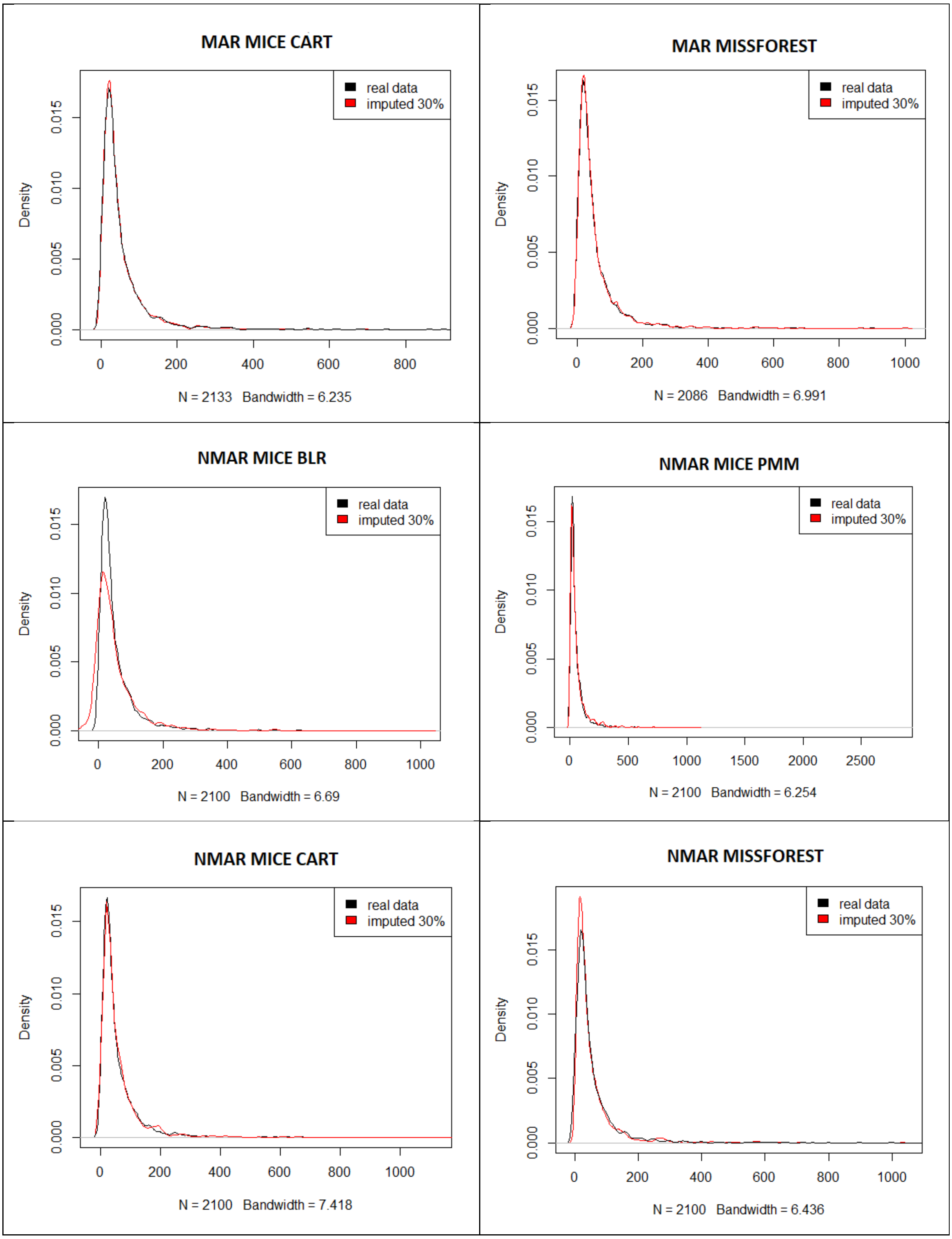
	MISSFOREST	MICE PMM	MICE CART	MICE BRL
men practising sports	0,9962	0,1846	0,9168	0,6451
women practising sports	0,9907	0,1899	0,9505	0,7070
male competitors	0,9771	0,1738	0,9628	0,6053
female competitors	0,9875	0,1918	0,9589	0,6725

ML methods outperformed non-ML methods in each case.

In the last stage of analysis of the results, some properties of distribution of imputed data were examined. Under MCAR assumption, mean of true values and mean of imputed values coincided in most of cases except MICE with PMM. Relative bias did not exceed 5%. Under MNAR, relative bias was higher in all cases, but with a little advantage of ML methods. Analysis of standard deviation lead to similar conclusions.

Below there are the density plots for real and imputed data for each method when the percentage of missing data was 30% (variable D2W3).





In a case when missing data pattern was MCAR, all methods achieved quite similar distribution, except MICE with PMM which produced some outliers. However, when missing data pattern was MNAR, distribution got worse for every method, but ML methods produced distribution more similar to distribution of real data than non-ML methods.

5. Code/programming language [e.g. Python, R; you can share your code here as a snippet, as separate file attachment or via Github, google Colab (see examples here)]

R language. Code and input data are stored on local servers.

6. Evolution of this study inside the organisation[e.g. Has this study advanced ML within the organisation?Was there any collaboration within the organisation?]

It is the first time when multiply imputation is tested. Some employees have already took their firststeps in this topic during the training conducted in our office.

7. Is it a proof of concept or is it already used in production?[If it is a proof of concept: Was it successful? How will its results prospectively be used in the future?]

7.1 What is now doable which was not doable before?

MissForest method from "missForest" R-package allows to perform multiply imputation with a few lines of code. Definitively, even not-advanced R-users can pick up that imputation method in R. Moreover, the imputation results cover also some precision measures without need of manual calculations.

7.2 Is there already a roadmap/service journey available how to implement this?

Not yet.

7.3 Who are the stakeholders?

Statistics Poland

7.4 Fall Back

Current imputation method was challenged and need to be changed. Thus, there is nothing to go back to.

7.5 Robustness

MissForest method did not produce results out of range of true values. Non-ML methods produced outliers under MNAR assumption.

8. Conclusions and lessons learned[e.g. ML can be used for editing but one has to have the following points in mind ...]

Machine Learning methods provided more precise outputs than non- Machine Learning methods. The results of data imputation with MICE strongly depended on underlying single imputation method. If it was CART, the results were better, in general, then for non-ML methods, that is Predictive Mean Matching and Bayesian Linear Regression. ML methods produced admissible results without outliers, what was not a case for not-ML methods. Distributional properties of data imputed with ML-methods were also better. ML preserved inequalities well. The only disadvantage of ML-methods, especially MissForest, is computational complexity.

9. Potential organisation risk if ML solution not implemented

10. Has there been collaboration with other statistical organisations, universities, etc?

No.

11. Next steps

In our institution there is a need to develop a relevant knowledge and skills to understand the process of building and testing ML models.

12. References or additional resources