

Computing Topics on Multiple Imputation in Big Identifiable Data Using R: An Application to Educational Research

Maria Eugénia Ferrão[1,2(✉)] and Paula Prata[1,3]

[1] University of Beira Interior, Covilha, Portugal
meferrao@ubi.pt, pprata@di.ubi.pt
[2] Centre for Applied Mathematics and Economics (CEMAPRE), Lisbon, Portugal
[3] Instituto de Telecomunicações (IT), Covilha, Portugal

Abstract. This article shows how to conduct multiple imputation in big identifiable data for educational research purposes. The R statistical package and procedures to handle missing data applied for the purpose of this study were "BaylorEdPsych" and "mi". Firstly, we checked that every dataset rejected the null hypothesis for Missing Completely At Random (MCAR), using the function "LittleMCAR". Simulated and real data analyses were conducted. Results suggest that the improvement of the quality of imputation requires alternative methods to be developed.

Keywords: Multiple imputation · R programming · Big data · Education research

1 Introduction

Several authors mention having written on the use of big data for research purposes, which may be simultaneously an opportunity or a threat (Diggle 2015). If the research purposes include inferences from big data statistical modelling and the big data does not cover the entire target population or there is a selective mechanism that produces missing data that are not completely at random, the research itself may be compromised.

The concept of missing data is defined as "the difference between the data we planned to collect and what we have managed to collect" (Longford 2005, p. 13). In several educational research studies, missing data are due to item nonresponse, i.e. participants in a survey or test who do not give responses for every item or question administered (item missing). In some situations the expected participants in the survey or test do not appear (subject missing).

According to Schafer (1999, p. 1), when the incomplete cases comprise only a small fraction of all cases (say, five percent or less) then case deletion may be a perfectly reasonable solution to the missing-data problem. In multivariate settings where missing values occur on more than one variable, however, the incomplete cases

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11621, pp. 12–24, 2019.
https://doi.org/10.1007/978-3-030-24302-9_2

are often a substantial portion of the entire dataset. If so, deleting them may be inefficient, causing large amounts of information to be discarded. Moreover, omitting them from the analysis will tend to introduce bias, to the extent that the incompletely observed cases differ systematically from the completely observed ones.

Schaffer (1999) proposed 5% as the cutoff, but other authors suggested 10% or even 20%. Moreover, research papers in education often do not mention the occurrence of missing data despite best practice recommendations in reporting and handling missing data (Pampaka et al. 2016; Schlomer et al. 2010) in quantitative based research.

In addition, the use of quantitative methods in the field of education and other related disciplines such as economics, sociology, political science, has been strongly conditioned either by the availability of data or by their quality (c.f. Bratti et al. 2004; Foley and Goldstein 2012). For example, an increasing number of researchers use large and complex data surveys, such as TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), TALIS (Teaching and Learning International Survey), or PISA (Programme for International Student Assessment) collected in more than 60 countries, or assessment and evaluation data collected within each country, for innovative contributions in the field of economics of education, sociology of education, psychology, etc. For certain research purposes, however, omit the occurrence of missing data or assume naïve solutions for missing imputation, could undermine the scientific innovative contribution.

Three main contributions arise from this paper: (1) What should the researcher do with incomplete data? (2) How to identify the pattern of missing data? (3) How to conduct multiple imputation with big identifiable data? The remainder of this paper consists of five parts. Section 2 presents concepts and definitions on missing data classification, and the explanation of a statistical test to identify the pattern of item missingness. Section 3 describes the R statistical package and procedures to handle item missing data. Section 4 deals with multiple imputation procedures applied to simulated and real data. The "Prova Brazil" 2017 (INEP 2018) was used as an example of educational research where identifiable big data (Shlomo and Goldstein 2015) does not cover the entire target population and, thus, the missing items may not be completely at random. The section includes data description, the simulation study and performance evaluation. Finally, the conclusion constitutes the Sect. 5.

2 Concepts and Definitions on Missing Data

Regarding data quality for statistical modelling purposes, the kind of missing data randomness matters (Little 1988; Little and Rubin 2002), i.e., it may occur as Missing Completely At Random (MCAR), as Missing At Random (MAR) or as Not Missing At Random (NMAR). "One way of assessing this is to compare the means of recorded values of each variable between groups defined by whether other variables in the dataset are missing or not" (Little 1988, p. 1198). The respective definitions and examples in education research are as follows (Little 1988). Consider $\mathbf{y}$ as the $n \times p$

data matrix where n is the number of subjects and p the number of variables; and **r** as the respective n × p missingness indicator matrix, such that,

$$\text{if } \ y_{ij} \text{ is missing } \ \text{then } r_{ij} = 1 \text{ else } r_{ij} = 0.$$

Consider also the set of unknown parameters Θ, Ψ such as the data and missing mechanism are specified by the following distributions:

$$f(y|\Theta) \text{ for } y \text{ and } f(r|y, \Psi) \text{ for } r.$$

Following the author, we write $y = (y_{obs}, y_{mis})$, where y_{obs} denotes the observed values of y, and y_{mis} denotes the missing values of **y**.

Rubin (1976) (cited by Little 1988) defined the missing occurrence as completely at random (MCAR) if

$$f(r|y_{obs}, y_{mis}, \Psi) = f(r|\Psi), \forall_{y_{obs}, y_{mis}},$$

which is the highest level of randomness, i.e., missingness does not depend on the observed nor on the missing values of y. The probability of an instance ij having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method may be used without risk of introducing bias on the analysis and inferences conducted.

Rubin (1976) (cited by Little, 1988) also defined MAR as

$$f(r|y_{obs}, y_{mis}, \Psi) = f(r|y_{obs}, \Psi), \forall_{y_{mis}},$$

meaning that missingness does not depend on y_{mis} but may depend on y_{obs}. Under such conditions, the author called the missing data mechanism as ignorable for likelihood-based inferences. The same is to say that the probability of an instance ij having a missing value for an attribute may depend on the known values, but not on the missing value itself. The remaining situation is classified as NMAR that occurs when the probability of an instance ij having a missing value for a variable that could depend on the value of such variable.

Big data is characterized by large volume of data available at high velocity and in varying formats. The review by Wamba et al. (2015), citing many other authors, adds veracity and value to the general principles of big data. The report of the American Association for Public Opinion Research (Japec et al. 2015) includes administrative data in the sources of big data. A definition for big identifiable data is presented in the special issue by Shlomo and Goldstein (2015), "If the elements in a data set can be meaningfully associated with a unit at a given place and time, such as an individual, institution, product or geographical location, then big data can be made fit for purpose for statistical inference." (p. 787). This is the conceptual and methodological framework we adopt in this study to investigate the topic of missing data imputation on the Brazilian open access big identifiable data.

3 Statistical Packages and Procedures

The R statistical package and procedures to handle missing data applied for the purpose of this study were "BaylorEdPsych" and "mi". Firstly, we checked that every dataset rejected the null hypothesis for MCAR, using the function "LittleMCAR" (Beaujean 2015). That function belongs to the "BaylorEdPsych" package (R Package for Baylor University Educational Psychology Quantitative Courses). Then, a process composed of four main steps was conducted, in order to get the respective complete data,

1. Missing values imputation;
2. Testing convergence;
3. If it does not converge then go to step (1);
4. Collecting results;
5. In the simulation study, assessing the error (root mean square error and the mean absolute error);
6. Adjusting the linear regression models;
7. Evaluating the results obtained.

To do multiple imputation, the "mi" package (Missing Data Imputation and Model Checking) was used. That package imputes missing values in an approximate Bayesian framework (Su et al. 2015). After converting the initial data into a missing_data.frame object, the main features of each "missing variable" can be observed. For instance, each "missing variable" is classified with a type (continuous, binary, etc.) and a family (Gaussian, binomial, etc.). That initial classification done by the "missing_data.frame" function can be modified by the user accordingly to his knowledge, using the "change" function. At that moment the "mi" function can be called to perform the imputation. An iterative algorithm is used to model each "missing variable" as a function of all the other "missing variables" and their missing patterns. Multiple chains are run and convergence is assessed after a pre-specified number of iterations. The number of chains used should be related with the number of cores available so they can be run in parallel. In this work both approaches, sequential and parallel, were tested. As the study was done in windows operating system, parallel processing is achieved via sockets. A cluster of n nodes was created, being n the number of available logical cores. Using sockets has higher overhead than parallelism via forking, but that can just be used in POSIX systems (Errickson 2017). After the conclusion of a number of iterations (30 by default) the convergence was tested with the "mipply" and "Rhats" functions. With "mipply", a function can be applied to an object of class "mi", that is, the object returned by the imputation process. That object contains for each chain the data evolution throughout all the iterations. Computing the "mean" of each complete variable in each chain, the user can stop the iterative process when the "means" are approximately the same. The "Rhats" function will monitor the "mean" and "standard deviation" of each variable over the iterations, across the chains. Values close to 1 mean convergence.

Both the root mean square error (RMSE) and the mean absolute error (MAE) are usually employed in performance evaluation studies. Some authors argue in favor of MAE (Willmott and Matsuura 2005), suggesting that the RMSE is not a good indicator of average model performance and might be a misleading indicator of average error,

others (Chai and Draxler 2014) contend that the proposed avoidance of RMSE and the use of MAE is not the solution to the problem, demonstrating that the RMSE is not ambiguous in its meaning, and that the RMSE is more appropriate to represent model performance than the MAE when the error distribution is expected to be Gaussian, which is the case in our study. We decided to use both metrics. To simplify, we assume that we already have n samples of model errors ε representing the error between the complete and the imputed data, calculated as $e_i, i = 1, 2, \ldots, n$. The RMSE and MAE statistics are calculated for each data set as

$$\text{MAE} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}}|\text{e}_\text{i}|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

In addition, a linear regression model was considered with Y as the response variable and X1, X2 as the covariates, in order to assess the impact of multiple imputation on the estimates of coefficients. A regression equation of the form

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \tau_i$$

explains the value of a dependent variable y_i in terms of a set a two observable variables $x1_i$, $x2_i$ and an unobservable random variable τ_i. According to the assumptions, the elements τ_i are distributed independently and identically with expected values zero and a common variance of σ^2. More details can be found in Greene (2011) for example.

4 Big Identifiable Data with Multiple Imputation

4.1 "Prova Brazil" 2017 Data

The education data used in this study is the "Avaliação Nacional do Rendimento Escolar" (INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2015), well known as "Prova Brasil". The "Prova Brasil" was created in 2005 under the scope of the Basic Education Evaluation System (SAEB) with the aim of assessing students learning at Brazilian public schools. It is a quasi census type applied to students at the 5th and 9th grades of primary education in schools with 20 or more students enrolled in these grades. It covers all Brazilian territory and is carried out every two years by the INEP, which is responsible for developing and applying educational assessments, and also the census of education. The "Prova Brasil" comprises standardized tests on Portuguese Language (reading) and Mathematics, as well as questionnaires targeting students, teachers, principals and schools. Tests are prepared following SAEB's Reference Matrix, which contains the description of competencies and abilities that guide the production of test items. The scores of student proficiency are obtained by item response models, allowing the use of Balanced Incomplete Block

Design (BIBD) and equating. SAEB's proficiency scales range from 0 to 500 with mean 250 and standard deviation 50 to 9th grade students.

For the purpose of this paper, we consider the 5th year data collected in 2017. The finite population is, then, a large identifiable sample of size N = 2.594 million students of its super population of the 5th grade. The performance scores in Maths and Reading are available for 2.170 (84%) million students and the socioeconomic status for 2.132 (82%) students (Alves and Ferrão 2018; Ferrão and Alves 2019).

4.2 A Simulation Study

In this section we carry out simulation studies to assess the impact of multiple imputation on estimates of linear regression given a non-ignorable pattern of missingness. The simulation study is based on samples drawn from "Prova Brazil 2017" using the Rondónia Federation Unit (FU = 11), involving 27874 students. For the purpose of simulations the complete data was generated by applying listwise to the real data. We considered three relevant variables for research in educational evaluation or educational equity: Student's performance in reading (Y), Student' socioeconomic status (X1) and Student's trajectory without grade repetition (X2).

Based on the complete data (n = 20408), we carried out four samples Y, X1, X2 with 5% (Miss5), 10% (Miss10), 15% (Miss15), and 20% (Miss20) of missing occurring randomly in the variables Y and X1, respectively for the group of low-performing students and poorer students. Descriptive statistics for complete and missing data are presented in Tables 1 and 2, respectively. We can observe that, as expected, the mean and median increase as the percentage of missing increases due to the generator process of missing values.

Table 1. Descriptive statistics of complete data.

	Y	X1	X2
Mean	0.073	5.082	0.775
Median	0.040	4.960	1.000
Variance	0.822	1.105	0.175
Skewness	0.172	0.525	−1.315
Kurtose	−0.313	0.510	−0.271

Table 2. Descriptive statistics of missing data.

	Y Miss5	X1 Miss5	Y Miss10	X1 Miss10	Y Miss15	X1 Miss15	Y Miss20	X1 Miss20
Mean	0.166	5.166	0.232	5.230	0.295	5.289	0.340	5.327
Median	0.100	5.000	0.200	5.100	0.300	5.200	0.300	5.200
Variance	0.747	1.022	0.723	0.995	0.721	1.003	0.742	1.028
Skewness	0.238	0.612	0.175	0.591	0.066	0.495	−0.027	0.425
Kurtose	−0.241	0.638	−0.134	0.689	−0.050	0.682	−0.079	0.551

Checking the Missing Pattern. As said before we started to assess each multivariate dataset for missing completely at random with the "LittleMCAR" function that uses Little's test. We tested the datasets Miss5, Miss10, Miss15 and Miss20 described above and four additional cases where the generated missing values are MCAR. MCAR10X1 represents a dataset with 10% of missing values randomly occurring in variable X1, MCAR10X1Y has 10% of missing randomly values occurring in variables X1 and Y. MCAR20X1 and MCAR20X1Y have the same structure but with 20% of missing values. Table 3 contains the results obtained.

Table 3. Results of Little's test.

	Chi-square	Degrees of freedom	p-value	Missing patterns	X1, number of missing values	Y, number of missing values
Miss5	594.50	5	<0.001	4	969 (4.75%)	983 (4.82%)
Miss10	733.52	5	<0.001	4	1979 (9.70%)	1970 (9.65%)
Miss15	929.28	5	<0.001	4	3115 (15.26%)	3083 (15.11%)
Miss20	911.31	5	<0.001	4	4036 (19.78%)	4078 (19.98%)
MCAR10X1	0.82	2	0.66	2	2107 (10.32%)	0
MCAR10X1Y	7.79	5	0.17	4	2107 (10.32%)	2115 (10.36%)
MCAR20X1	2.91	2	0.23	2	4096 (20.07%)	0
MCAR20X1Y	3.53	5	0.62	4	4096 (20.07%)	4065 (19.92%)

No matter the percentage of missing data, the results show that the Little's test rightly identifies the pattern of non-ignorable missing data at the level of significance of 5%, since p value <0.001, as can be observed in Table 3. In addition, the test pro-vides p values greater than 0.05 for datasets where missing were generated as completely at random.

Results of Simulation Study. Table 4 contains descriptive statistics of imputed data. It can be observed that the values of mean and median tend to increase as the multiple imputation was applied to larger percentage of missing values, and the dispersion tends to be reduced.

Table 4. Descriptive statistics of imputed data.

	Y MI5	X1 MI5	Y MI10	X1 MI10	Y MI15	X1 MI15	Y MI20	X1 MI20
Mean	0.143	5.162	0.186	5.224	0.238	5.280	0.266	5.317
Median	0.100	5.000	0.100	5.100	0.200	5.200	0.300	5.200
Variance	0.751	1.020	0.732	0.997	0.728	1.010	0.748	1.030
Skewness	0.228	0.590	0.154	0.532	0.053	0.421	−0.030	0.335
Kurtose	−0.199	0.611	−0.088	0.625	−0.029	0.588	−0.040	0.489

The Figs. 1 and 2 illustrate the RMSE and MAE for different values of missing percentage. The effect of increasing the percentage of missing data does not change

substantively the relationship between the variables, but the magnitude is lower than it should be. It can be observed that the error is higher in X1 than in Y.

The Table 5 includes the correlation between Y and the linear predictor (R), the coefficients estimates (B1, B2) and the respective standard errors (SE(B1), SE(B2)), for regression models applied to the complete data (1st row), to the datasets resulting from multiple imputation (2nd to 5th rows) and to the respective data considering the missing values (6th to 9th rows).

Considering the results of the model applied to complete data (with R = 0.369) and comparing them with those obtained from multiple imputation, it can be observed that the maximum absolute difference on correlation is 0.015 (MI5 data, R = 0.354). However, differences on parameter estimates are larger in MI15 for B1. Concerning the standard errors, the results are always lower or equal than those obtained for complete data. Regarding the B1 estimates obtained either by multiple imputation or with the assumption of MCAR, the results are generally biased towards zero. It should be noted that the correlation is lower when MCAR assumption is used no matter the percentage of missing data. Since X2 is a complete data variable, a clear pattern for B2 estimates was not found.

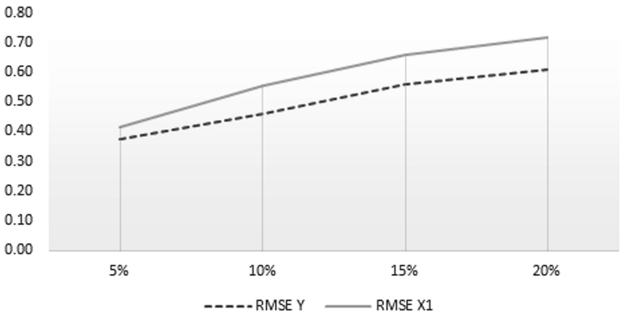

Fig. 1. Root square mean error.

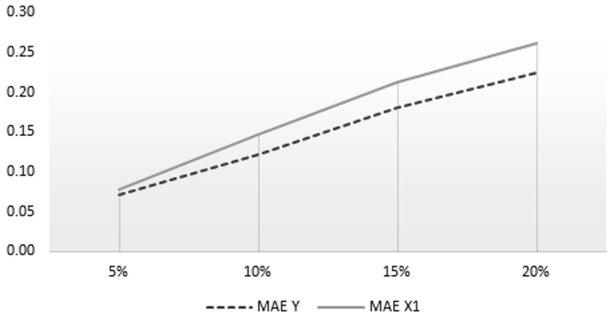

Fig. 2. Mean absolute error.

Table 5. Estimates of the linear regression models.

	R	B1	SE (B1)	B2	SE (B2)
Complete data	0.369	0.135	0.006	0.688	0.014
MI5	0.354	0.126	0.006	0.638	0.014
MI10	0.359	0.121	0.006	0.649	0.013
MI15	0.359	0.113	0.006	0.656	0.013
MI20	0.365	0.114	0.006	0.679	0.014
Miss5	0.347	0.127	0.006	0.639	0.015
Miss10	0.350	0.122	0.006	0.658	0.016
Miss15	0.340	0.122	0.007	0.651	0.017
Miss20	0.350	0.121	0.007	0.696	0.018

Figures 3, 4, 5 and 6 show the plots of the imputation process for datasets Miss5 and Miss20. Figure 3 shows on the left the histogram of completed data (in grey), the histogram of observed data (blue line) and the histogram of imputed data (red line) for the standardized variable X1 of Miss5 dataset. The plot in the middle of the same Figure shows a comparison of completed data (in blue) to the fitted values (in red). Colours just can be observed in the online version. Finally on the right hand side of the Figure the associated binned residuals are presented. Figure 4 represents with the same structure the values for the standardized variable Y of Miss5 dataset. Figure 5 shows the plots for the standardized variable X1 of Miss20 and Fig. 6 shows the plots for the standardized variable Y of Miss20 dataset.

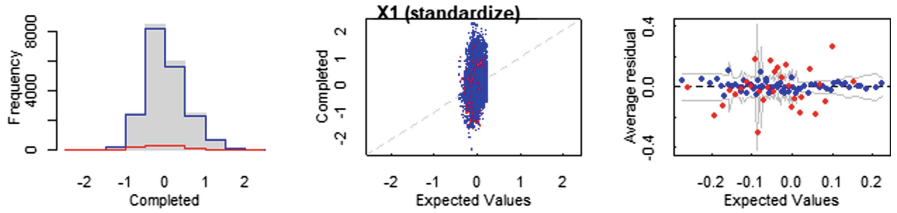

Fig. 3. Histogram of observed, imputed and completed data (at the left hand side), comparison of completed data to the fitted values (at the middle) and the associated binned residuals (at the right hand side) for variable X1 (standardized) of Miss5 dataset. (Color figure online)

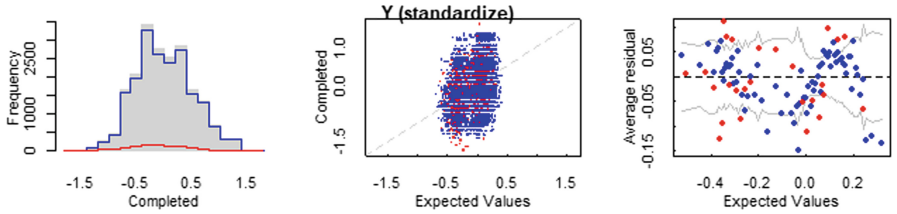

Fig. 4. Histogram of observed, inputted and completed data (left hand side), comparison of completed data to the fitted values (at the middle) and the associated binned residuals (left hand side) for variable Y (standardized) of Miss5 dataset. (Color figure online)

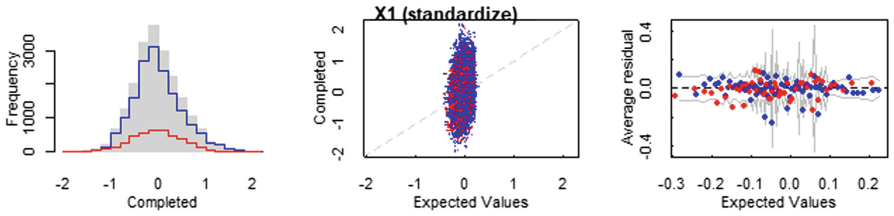

Fig. 5. Histogram of observed, inputted and completed data (left hand side), comparison of completed data to the fitted values (middle) and the associated binned residuals (right hand side) for variable X1 (standardized) of Miss20 dataset. (Color figure online)

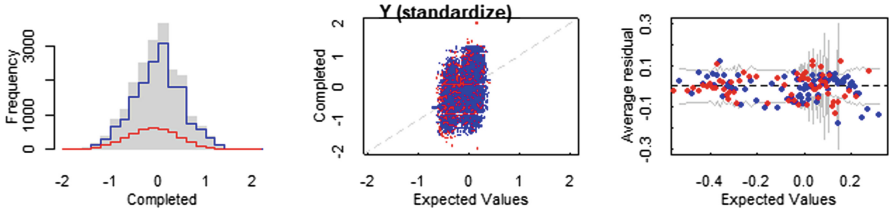

Fig. 6. Histogram of observed, inputted and completed data (left hand side), comparison of completed data to the fitted values (middle) and the associated binned residuals (right hand side) for variable Y (standardized) of Miss20 dataset. (Color figure online)

The graphs of Figs. 3, 4, 5 and 6 confirm that the imputation on Y presents a smaller error than on X1. Despite the impact of the error on regression model estimates being always lower than 0.02, the multiple imputation can be improved by increasing the number of variables included in the joint distribution.

4.3 Execution Time and Performance

This section describes the limitations imposed to this study by using a common laptop. We analyze the execution times needed to run multiple imputations when varying the number of records present in the data files, the number of variables and the type of variables.

The execution times were obtained in two computers. A laptop that will be referred as LT1, with an I7-4600U processor (speed from 2.1 GHz up to 2.7 GHz) with 2 physical cores (four threads) and 8 GB of RAM. A second laptop referred as LT2, is a machine with an I7-8550U processor (speed from 1.8 GHz up to 4.0 GHz, with 4 physical cores (8 threads) and 16 GB of RAM. Both computers run a Windows 10 operating system (64 bits version). Besides the processing speed, the biggest difference between the two machines is the available memory. We got the time at the start and at the end of the block of code in study with the function Sys.time() and estimated the time after running 3 times the same code.

Starting by studying the impact of using the "mi" function in parallel in a Windows system we got the results presented in Table 6. It shows the approximations for parallel

and sequential execution times, using machine LP1, when multiple imputation is done in a missing_data.frame with N observations on 3 continuous variables. In all cases, 2 of the variables have approximately 5% of missing values and there are 4 missing data patterns. Three FUs were chosen considering the number of observations in the studied variables: a small one, Rondônia (FU = 11) a medium one, Ceará (FU = 23) and a large one, São Paulo (FU = 35). In all the experiments 30 iterations were done and the imputation process converged. As can be observed from Table 6, parallel execution, in cases with a small number of missing patterns, just has advantage for data with more than 20 000 observations. For smaller cases the overhead of creating a cluster surpasses the gain of running the different chains in parallel.

The experiments on datasets with different percentages of missing values, as Miss5, Miss10, Miss15 and Miss20 used in our simulation study, showed small variations in execution times. Thus we conclude that for datasets of about 20 000 observations and few variables the execution time does not depend on the missing percentage.

Table 6. Execution times of multiple imputation, sequential versus parallel when varying the number of observations.

Observations	Missing patterns	Variables with missing values	Execution time (Parallel)	Execution time (Sequential)
20408 (FU – 11)	4	2	30 to 40 s	40 to 50 s
86628 (FU – 23)	4	2	2 to 3 min	3 to 4 min
376325 (FU – 35)	4	2	4 to 5 min	14 to 15 min

Next, we increased the number of observations, increased the number of variables with missing values and included binary variables. All studied cases have more than half a million observations (N = 513304). The execution times obtained in both machines, LT1 (8 GB of RAM) and LT2 (16 GB of RAM) are shown in Table 7.

Table 7. Execution times of multiple imputation for datasets with 513304 observations varying the number and type of missing variables in a computer with 8 GB of RAM versus a computer with 16 GB of RAM.

Missing patterns	No. of variables (with missing)	Class of variables with missing	Execution time (8 GB of RAM)	Execution time (16 GB of RAM)
8	6 (4)	Continuous	29.9 min	9.3 min
8	6 (4)	2 continuous 2 binary	33.0 min	9.6 min
18	10 (8)	Continuous	116.4 min	30.8 min

The case presented in the first row, has 6 continuous variables being 4 with missing values. For that case the execution times vary from about 10 min in LP2 to almost 30 min with LP1. In the second case two of the variables with missing values are binary. As can be seen in that case the execution time is slightly larger in both machines but the difference is inconclusive. In all cases the computer memory is fundamental to

the performance of the imputation routine; this is especially visible in the last case when we double the number of variables with missing values. With 18 data patterns the limit of memory in LT1 machine is almost reached. Finally, it should be noted that those three cases were executed running 4 chains in parallel, with a limit of 35 iterations, but some variables did not converge. It would be necessary to continue iterating.

5 Conclusion

In this article we showed how to use R packages to conduct multiple imputation in big identifiable data for the purpose of educational research. As long as the number of cases and the number of variables increase, the time also increases. Preliminary results suggest that the improvement of multiple imputation via the inclusion of additional variables is limited to execution time and the machine processor and memory available. Thus, as future work we intend to explore the use of a distributed computing engine like Apache Spark either in machines with more memory and cores or in cloud platforms. Further research to include categorical variables in the process of multiple imputation is going on, and in order to validate the conclusions a Monte Carlo study is planned to be conducted.

Acknowledgements. This work was partially funded by FCT- Fundação para a Ciência e a Tecnologia through project number CEMAPRE - UID/MULTI/00491/2019 and by FCT/MEC through national funds and when applicable co-funded by FEDER – PT2020 partnership agreement under the project UID/EEA/50008/2019.

References

Alves, M.T.G., Ferrão, M.E.: Uma década de Prova Brasil: Evolução do desempenho e da aprovação (2018, submitted)

Beaujean, A.A.: Package "BaylorEdPsych" (2015). https://cran.r-project.org/web/packages/BaylorEdPsych/BaylorEdPsych.pdf. Accessed 11 Feb 2019

Bratti, M., McKnight, A., Naylor, R., Smith, J.: Higher education outcomes, graduate employment and university performance indicators. J. R. Statist. Soc. A **167**(3), 475–496 (2004). http://www.jstor.org/stable/3559775

Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. Model. Dev. Discuss. **7**(1), 1525–1534 (2014)

Diggle, P.J.: Statistics: a data science for the 21st century. J. R. Stat. Soc. A **178**, 793–813 (2015)

Errickson, J.: Parallel processing in R (2017). http://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/parallel.html. Accessed 10 Feb 2019

Ferrão, M.E., Alves, M.T.G.: Grade repetition in Brazilian primary education: 2007–2017 cross-sectional data modelling (2019, submitted)

Foley, B., Goldstein, H.: Measuring Success: League Tables in the Public Sector. British Academy, London (2012)

Greene, W.H.: Econometric Analysis. Prentice Hall, New York (2011)

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, ANRESC (Prova Brasil) (2015). http://portal.inep.gov.br/educacao-basica/saeb/sobre-a-anresc

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C.: AAPOR Report on Big Data (2015). https://www.aapor.org/getattachment/Education-Resources/Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.aspx

Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. **83**(404), 1198–1202 (1988)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley, Hoboken (2002)

Longford, N.T.: Missing Data and Small-Area Estimation. Springer, New York (2005). https://doi.org/10.1007/1-84628-195-4

Pampaka, M., Hutcheson, G., Williams, J.: Handling missing data: analysis of a challenging data set using multiple imputation. Int. J. Res. Method Educ. **39**(1), 19–37 (2016). https://doi.org/10.1080/1743727X.2014.979146

Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, Boca Raton (1999)

Schlomer, G.L., Bauman, S., Card, N.A.: Best practices for missing data management in counseling psychology. J. Couns. Psychol. **57**(1), 1–10 (2010). https://doi.org/10.1037/a0018082

Shlomo, N., Goldstein, H.: Editorial: big data in social research. J. R. Stat. Soc. A **178**, 787–790 (2015)

Su, Y.-S., Goodrich, B., Kropko, J.: Package "mi" (2015). https://cran.r-project.org/web/packages/mi/mi.pdf. Accessed 11 Feb 2019

Wamba, S.F., Akter, S., Edwards, A., Chopin, G., Gnanzou, D.: How "big data" can make big impact: findings from a systematic review and a longitudinal case study. Int. J. Prod. Econ. **165**, 234–246 (2015). https://doi.org/10.1016/j.ijpe.2014.12.031

Willmott, C., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82 (2005)