Performance of Imputation Methods towards Increasing Percentage of Missing Values

Kenfac Dongmezo Paul Brice, Peter N. Mwita
Machakos University
Kamga Tchwaket Ignace Roger
Pan African University Institute for Basic Sciences,
Technology and Innovation (Pauisti), Kenya
Email – dongmezobrice@gmail.com /
brice.dongmezo@students.jkuat.ac.ke
Email – petermwita@mksu.ac.ke

Abstract

The aim of this paper is to study the performance of eightdifferent existing imputation methodsused on simulatedand real dataset. The methods are compared in term of their ability to estimate the missing observationsand estimate some statistics(mean, standard deviation and coefficient of a regression) using the full data set completed by the imputation. The comparisons are made using root mean square error, mean absolute deviationand bias observed after estimation of statistics. Simulation results using specific simulated data and bootstrap show that Mean Imputation and Complete case analysisare the best method in completing the data set and in obtaining best estimators for statistics.However, the results are subject to major changes if parameters like sample size, number of replication and type of distribution chosen are modified. In short with real data, result will change depending on the structure of dataset to impute. For example, application of the simulation results to a Rwandan dataset on smallholder farmers revealed that k-NN is the best method in reconstructing and Multiple Imputation can be used as imputation method in case we are to estimate some statistics. Our final conclusion is that imputation methods cannot be compared since in most cases their performance is parametrically linked to the data. We finally proposed a methodology and a simulation protocol to identify for any data set which imputationmethod will give the best results and therefore should be applied in priority.

Key words. Bias, Bootstrap, Imputation,Root Mean Squared Error, Mean Absolute error.

Introduction

Missing data is a common problem in applied statistics when dealing with collected data. It is a classical problem in all areas of research including: biology (Troyanskaya O et al, 2001), medicine

(Lewis HD, 2010), climatic science (Schneider T, 2001) and others. Nearly all standard statistical methods presume complete information for all the variables included in the analysis. However, a relatively few missing observations on some variables can dramatically shrink the sample size and affect the quality of estimators produced from those data (Marina Soley-Bori, 2013). After data collection where sampling has been done properly, often the data set will come with blank spaces meaning that some questions have not been responded to during survey or some specific information were not collected properly. This situation raises one main question: how can we manage the units with the missing information?

Many researchers have proposed steps to study the problem of missing data, starting by the missingness mechanism, why some observations are missing? Then follows the decision between dropping units with missing observation or imputation. Finally, in case imputation is chosen, which imputation method to adopt considering the situation.

This paper aims to analyse the performance of imputation methods toward an increasing percentage of missing values and draw the related conclusion on comparing imputation methods. The paper is organized as follows: Section 2 discusses the missingness mechanism with some typical examples and implication of having missing data in the set. Section three presents some of the most used imputation methods including the most recent like Multiple imputation and weighting. Sections 4 and 5 investigate and discuss the simulation and the results obtained from simulation. Finally, section 6 concludes and introduces different uses of imputation methods beyond replacing missing data.

1. Missingness Mechanism

Early works on missing data were carried out by Rubin (1987, 1996). Close to that, some researcher like Afifi and Elashoff (1966), Hartley and Hocking (1971), Ochard and Woodbury (1972) andLittle (1971) did a bit more on the topic with some applications in different areas of study. Most of these works started with the missingness mechanism.

Prior to presentation of general imputation methods or how to handle missing data problems, it is good to know why these data are missing. We present different missing data mechanisms, meaning how in our data base missing values appeared? There are 4 main situations where data can be missing:

- ✓ Missingness completely at random (MCAR): the probability of missingness is the same for all unit in the sample. For a given variable X in the data base, the probability for an observation to be missing does not depend on X itself and on other variables of the same data base. The perfect example will be if the choice is given to respondent to answer a question or not given a random condition (rolling a dice for example). It is difficult to have this situation in the real survey but it is the most common hypothesis in simulation or with real data imputation (Briggs et al., 2003; Allison, 2001).
- ✓ Missingness at random (MAR): Most missingness is not completely at random, as can be seen from the data themselves. Probability can depend on an auxiliary variable in the same survey. Respondent can decide to answer or not, or interviewer may forget to ask some questions to respondents. A more general assumption, missing at random, is that the probability a variable is missing depends only on available information. Thus, if for example sex, race, education, and age are recorded for all the people in the survey, then "earnings" is missing at random if the probability of nonresponse to this question depends only on these other fully recorded variables (Allison, 2001; Gelman& Hill, 2006).
- ✓ Missingness that depends on unobserved predictors (NMAR): Missingness is no longer "at random" if it depends on information that has not been recorded and this information also predicts the missing values. There are some underlying unobserved factors that could lead people not to answer a given question and they can differ from one person to another. Therefore, the probability of missingness is different across unit in our survey. An example is when during a survey a corrupted person is not going to declare his revenue because he knows that if he declares he can be exposed to pursuit because of corruption (information not recorded) the data will be missing (Allison, 2001; Gelman & Hill, 2006).
- ✓ Missingness that depends on the missing value itself: Finally, a particularly difficult situation arises when the probability of missingness depends on the (potentially missing)

variable itself. For example, this often happens because people are unlikely to reveal a high income to avoid being exposed (Allison, 2001; Gelman & Hill, 2006).

All these types of missingness can happen during survey and can be observed in data set depending on variables and the data collection process. To identify the type of missingness, the final data set ultimate user should be close to the data base constructor or be involved in data collection. The most frequent type of missingness mechanism is MAR. Practically, this is the one which can easily happen.

2. Different Imputations methods

In handling, missing data, we have two possibilities: discard missing data or imputation. Discard some unit presenting missing cases implies to reduce significantly the sample size especially in case more than one variable present missing data at different lines (cases). As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased (Soley, 2013). So, the best solution will be to impute data. There are several direct and simple methods of imputation including: Mean imputation (replace missing values by the mean or conditional mean or marginal mean of the variable), Last value carried forward (use the last value from a unit which logically is supposed to be close to the missing one), Using information from related observations (impute by a value from an individual which is closed to the missing one), Indicator variables for missingness of categorical predictors (add an extra category for the variable indicating missingness.), Indicator variables for missingness of continuous predictors (replace the missing value by a zero or by the mean), Imputation based on logical rules (use the logic of questionnaire to impute a value) (Allison, 2000, 2003).

As we said earlier, our research focuses on comparing any action taken to deal with missing data including discarding cases with missing data. Classical imputation methods are divided into two main groups. Let's assume that our variable of interest with missing observation is Y and the set of covariates without any missing observation is X . To simplify notation, forget about the indexes specifying the case. A missing observation in the set is denoted by Y_m and a non-missing one by

Y_{nm}. Of course, the corresponding covariates will be X_m and X_{nm} but it doesn't mean that they are missing.

2.1 Imputation methods that doesn't incorporate random variation

The main characteristic of these methods is that the missing value is replaced by a single estimator of the true value. They are deterministic methods meaning that there is no randomness in the set of values used for imputation. Running the same method on the same sample will always produce the same imputed values for unit missing with the same characteristics.

Mean Imputation and Conditional Mean Imputation

This method can be applied on any type of dataset, with or without covariates. It recommends to replace the missing value by the mean of the missing variable obtained using the non-missing observations. The user can just replace the missing observations in Y by the marginal mean directly: Y_m □ $E(Y_{nm})$ or knowing some properties of Y, conditional mean can also be used. The mean of Y given certain existing covariates X in our dataset: $Y_m = E(Y_{nm} / X = x_m)$. For example, if among our covariates, there is a variable sex and our variable of interest is determined by sex, we divide our sample into two groups: male and female, then perform mean imputation in each group. It is the most used method even if it leads to biased estimates and low variance and covariances (generally underestimate variances).

Nearest Neighbours Imputation

To apply this method, a data set with a set of covariates is absolutely needed. The first step of this method is to define what is a neighbour using the set of covariates X. To define a neighbour, there is a need to define the distance between cases. The default distance is the Euclidian distance: $d_{ij}^2 = (X_i - X_j)'(X_i - X_j)$. We can also use the Mahalanobis distance by introducing a transfer matrix in the Euclidian distance. After defining a distance, the user can now decide for a given missing value which case is close to it or not. You can replace the missing observation by the value of the nearest neighbour or by a fixed k nearest neighbour (averaging) or use a value obtained by all the data set weighting each available case by the inverse of the distance between the missing case and all of them (weighting average). The simulation in this study used the Gower distance developed by Gower (1971) which aggregate all the distances between two points for each variable in one single quantity. The distance was included in the package VIM on R by Kowarik (2016).

Last value carried forward

This method recommends to use the last value known about the variable for imputation. It means that if we have another survey, collecting the same information a time before the actual survey, from that survey you take information from the same variable and impute to the missing value in the actual data set. This method assumes the value doesn't change much with time. It can be true for some variables like sex but it is not always true.

Regression to perform deterministic Imputation

The method is a model-based method. It uses econometric (linear regression model or quantile regression for example) to build a model with available cases of Y and their corresponding covariates. The deterministic part of that model is used to predict the missing values given that all the values of the covariates for each of them are known: $Y_m \square f(X_{nm})$. The main advantage of this method is the fact that it uses all information available on different units to predict the missing value and with a good R^2, imputation can give interesting results. The disadvantages are: it overestimates model fit and correlation estimates and weakens variance of the variable Y .

Simple random Imputation (Hot deck imputation)

This method recommends to randomly select a set of available cases among our non-missing observations and impute them to the missing observation or for each missing observation, randomly select another one among the set of observed data and impute, $Y_m \square Sample(1, Y_{nm})$. This method is quite simple and looks interesting but for some database and if you want to perform some specific studies, results can be very bad. It doesn't take into account the covariates if they are available, consequently you can have some atypical case for example a 12-year-old child with a PhD as educational level. This method is suitable if the population is stratified according to some determinant of our variable of interest.

2.2 Imputation methods that do incorporate random variation

This group of methods are characterized by the fact that it allows for randomness in the prediction of missing values. Running this method n times in a given sample may produce n different values for a single imputation. Some of the methods presented here can be repeated then the final imputed value will be the average of the different output obtained during repetition.

Regression to perform random Imputation

This imputation method is almost the same as regression presented in the previous section. It also uses suitable econometric models to build a function of covariates that are going to be used to predict the value of the missing observation. The difference now is the error. After estimation of the coefficients of the regression, we obtain the deterministic part of the model and the error. Knowing the distribution of the error, this method recommends to generate for each predicted value an error and add to the deterministic part to obtain the final predicted value. The result is of the form: $Y_m \square f(X_{nm}) + \varepsilon_m$, with ε_m following a specific distribution determined by the econometric model. The main advantage here is the fact that the variance of the variable is preserved due to the randomness of predicted values. The drawback is the same, estimation of coefficient comes with some bias because the coefficient that we are using in the model are not the true coefficients but just estimators which of course brings another bias.

Multiple Imputation (MI)

Among imputation methods, Multiple imputation is one the most interesting methods and most performant according to literature. The main objective of this method is to replace the full set of missing values by different sets of possible candidates provided (each set) by different methods or by a single method allowing random variation. Multiple Imputation is a simulation procedure and the aim is not to obtain imputed values close enough to the real one but obtain acceptable estimators from the completed dataset (Schafer, 1997).

Multiple imputation involves three main steps:

a) **For each missing observation, generate m imputed values to obtain m completed sets of data**. After identifying which variable has missing values, the user should identify the missingness pattern and then decide which imputation methods to use keeping in mind that each should allow for randomness;

b) **Analyse the m set of completed data using standard procedures to produce estimators that we want**. In our case, each completed data set will produce some estimators;

c) **All the estimators produced from each completed data set are combined to form a single set of final estimates of the parameters of interest.** In this step, the average can be used to obtain the final parameters with a standard deviation and confidence interval.

As advantage, this method can be used with any kind of data and model. It is simulation based therefore any user who is good in programming can perform it in using any software. When data is MAR, Multiple Imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates. The main drawback is instability of the method. Because of randomness, different users can perform it and obtain totally different results. Even the same user, every time you run the program, you obtain different results hopefully slightly different. In the simulations, the MI method used generates Multivariate Imputations by Chained Equations (MICE). In the MICE procedure, a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution, with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression.

Maximum likelihood Imputation (ML)

This method is used to obtain the variance-covariance matrix for the variable in the model based on all the available data points, and then use the obtained variance-covariance matrix to estimate the regression model (Schafer, 1997). This method is quite simple if you use an appropriate software, you only need to specify your model of interest and indicate that you want to use ML. Theoretically, the basic idea is as follows. Given a set of data with n independent observations and $k+1$ variables $(y_i, x_{i1}, ..., x_{ki})$ and assuming that there is no missing data in that set, the likelihood function is given by:

$$L = \prod_{i=1}^{n} f_i(y_i, x_{i1}, ..., x_{ki}; \theta) \qquad (3.2.1)$$

where $f_i(.)$ is the joint probability function of i observations and θ the set of parameters to be estimated. The ML estimates are the values of θ that maximise L. Now, in the specific case of this research, suppose that for some observations i, the first variable Y has missing data that satisfies MAR assumption of missingness. Now the joint probability of the observed data is given by:

$$f_i^*(x_{i1}, ..., x_{ki}; \theta) = \int_y f_i(y_i, x_{i1}, ..., x_{ki}; \theta) dy \qquad (3.2.2)$$

For each observation's contribution to the likelihood function, we integrate over the variables that have missing data, obtaining the marginal distribution of observing those variables that have actually been observed.

Considering that there are m missing observations in the first variable over n, ordered such that the first $n-m$ lines are completed and the last m have missing data, the likelihood function of the full data set becomes

$$L = \prod_{i=1}^{n-m} f_i(y_i, x_{i1}, ..., x_{ki}; \theta) \prod_{i=n-m+1}^{n} f_i^*(x_{i1}, ..., x_{ki}; \theta) \qquad (3.2.3)$$

This likelihood function can then be maximized to get ML estimates of θ using several different methods.

There are two main ML methods:

a) **Direct Maximum Likelihood**: implies direct maximization of the multivariate normal likelihood function for the assumed linear model.

b) **The expectation – Maximization (EM) algorithm**: provides estimates of the mean and covariance matrix which can be used to get consistent estimates of the parameters of interest.

For the simulation, the R package MissMech is chosen. Two options are used to perform ML: firstly, the program assumes that data follow a multivariate normal distribution then secondly no assumption is made on the distribution but a maximization algorithm is used to obtain the covariance matrix.

3. Simulations and Results

This section presents an analysis of performance of different imputation methods on a simulated data set. The aim is to answer the question: which imputation methods gives better results in terms of reconstructing dataset and in terms of leading to better estimates of some statistical quantities for simulated data?

3.1 Simulation protocol

To simplify our analysis, we assume that there is only one variable y with missing observations in the data set with in the sample of size n. In addition to that, there are some covariates x_1, x_2 and x_3 generated given specific distributions (continuous and discrete) which determine the variable y.

Initially, the variables y, x_1, x_2 and x_3 are generated without missing value according to the regression equation $\hat{y} = \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \hat{\alpha}_3 x_3$. With the data set without missing values (sample size n), we compute the true sample value of the mean μ of y, the standard deviation σ, the coefficients α_i already known, in short the vector $param = (\mu, \sigma, \alpha_1, \alpha_2)$ is computed.Then, we gradually create missing data in the data set for the variable y from 10% of missing values up to 60% with a step 10%, 6 different percentages of missing values. For each percentage of missing values generated, firstly the vector $param$ is estimated using the complete case available (listwise deletion). Secondly, using specific imputation methods, the s% missing is estimated and then the vector $param$ is again estimated in a bootstrap of 1000 replication and compared to the true value. In addition to $param$ in the second step, the RMSE and MAE are computed to see how good the imputation methods were.

Steps of simulation

- **Step1**:Generate a sample of n observation of the random vector (Y, X_1, X_2, X_3) such that there is a linear and significant link between Y and the X covariates: output $(Y_i, X_{1i}, X_{2i}, X_{3i})_{i=1}^{n}$.
- **Step 2**: Compute the population or the full sample parameters from the simulated data such that $param = (\mu = mean(Y), \sigma = std(Y), \alpha_1, \alpha_2)$; where α_1 and α_2 are coefficient of X_1 and X_2 in the linear regression $Y = f(X_1, X_2)$.
- **Step 3:** Create randomly s percent of missing value in the vector Y with $s \in \{10, 20, 30, 40, 50, 60\}$, leading to six Y variables with different percent of missing values.
- **Step 4:** For each percentage of missing value, first compute the vector $param$ using complete case analysis meaning cases with missing data are deleted before estimation. Secondly, using each imputation methods selected, impute the missing values and compute the vector $param$ and the quantities $RMSE$ and MAE.
- **Step 5:** Compare the output of the simulation in bootstrap procedure of 1000 replications. Firstly, compare the vector $param$ for complete case analysis and for the one obtained in each imputation method to the real value of parameters and for different percentage of missing values (to see which method is best in estimating the true parameters). Secondly, compare $RMSE$ and MAE for different imputation method and different percentage of missing values (to see which method is best in reconstructing data).

As said in the last step, to make sure that the results are robust and to get standard errors, the simulation is associated with a bootstrap procedure of 1000 replication (creation of s percent of missing value 1000 times).

3.2 Results and discussion

All the simulations were done with a sample size of 1000 unit and 1000 replication in the bootstrap (for a given percentage, sampling 1000 times missing values) to see stability of results. Here is summary of results from two points of view: Reconstruction of data and ability to give better estimates of the full sample parameters. The results are specifically for the simulated data that we have, changing parameters of simulation can lead to other results.

3.2.1 Ability to reconstruct the data

The general comment on the results is that the value of RMSE is almost the same for all percentages of missing value for a given imputation method, with a slight increase for higher percentages of missing values. Figure 1 shows that for ML imputation, the RMSE is around 109 for the first 3 percentages of missing values but slightly above 110 for the last 3. This remark is the same for all the 7 RMSE computed. In addition to that, the error observed on RMSE is quite smallmeaning that the results obtain after simulations are quite stable and are not due to randomness. Comparing now different imputation methods, Figure 1 shows us that the best imputation method in data reconstruction (smallest RMSE) is Mean Imputation no matter the percentage of missing value, with an average RMSE of 100.78, followed by Regression Imputation without randomness with an average RMSE of 101.12 among all the percentage of missing values.

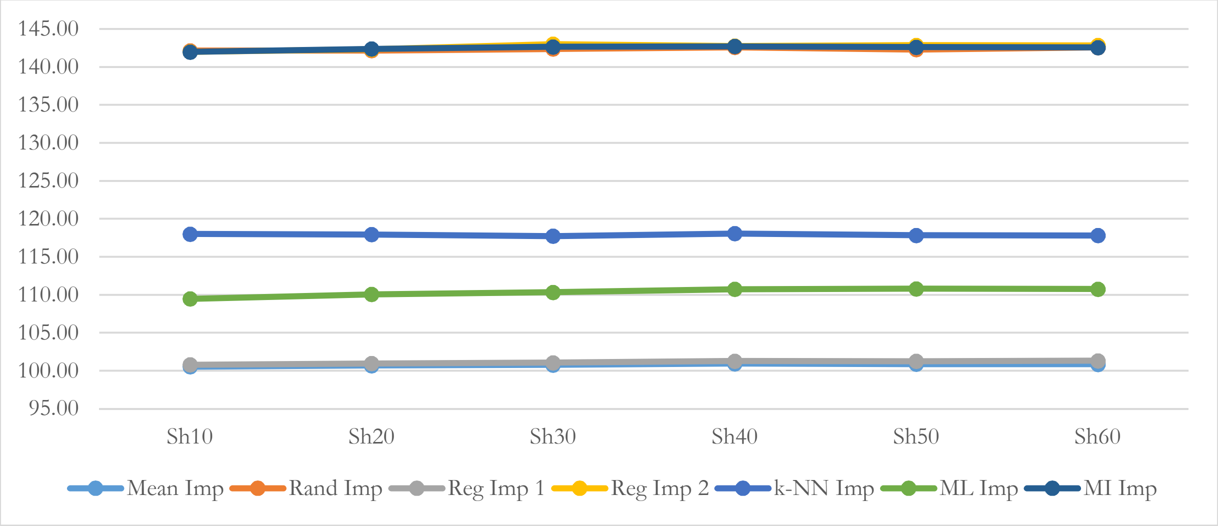

Figure 1: RMSE for imputation methods per percentage of missing value

The methods performing less than the others are Random imputation, Regression imputation with randomness and Random imputation. Their RMSE is above 140 which is clearly above all RMSE observed.

When we look at the MAE trends in Figure 2, the tendency is the same as for the RMSE. The value is quite constant along the different percentages of missing values but with a slight increase when the percentage increase. The errors are also small meaning a good stability in results.

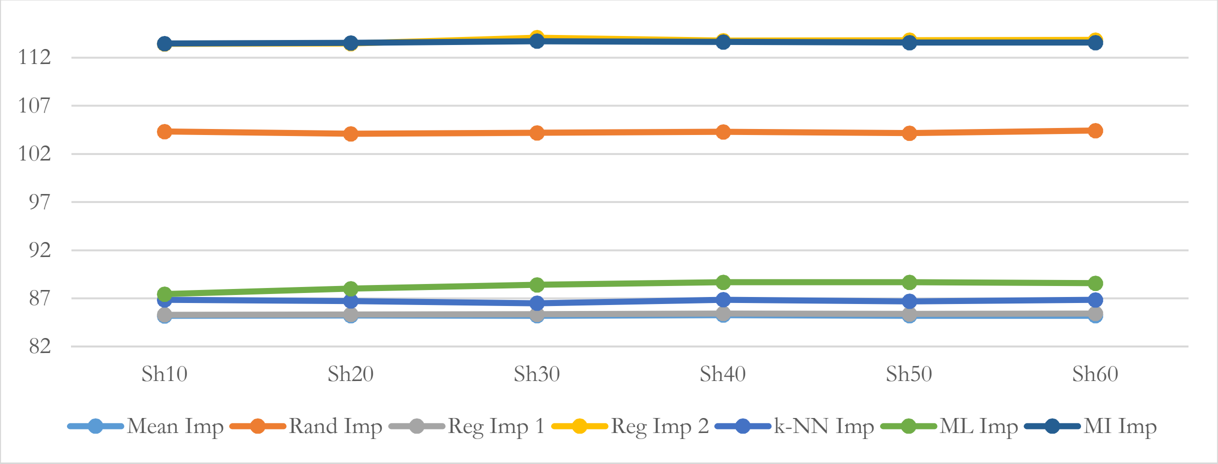

Figure 2: MAE for imputation methods per percentage of missing value

Here again the best imputation method is Mean imputation with an average MAE of 85.22 among all the percentages of missing value tested. The second best is the deterministic Regression Imputation with an average MAE of 85.38. When we look at the methods with the highest MAE, we find that Multiple imputation and Regression imputation with randomness are the one with the bad results.

To sum up, the best methods in data reconstruction are Mean imputation and deterministic Regression imputation. The weaker methods in data reconstruction are Multiple Imputation and Regression with Randomness. It seems like in general, with simulated data (given distribution clearly known) imputation methods including randomness performs less compared to those without randomness.

3.2.2 Ability to estimate full sample parameters

Secondly, the methods are tested on their ability to estimate some statistics computed on the full sample without missing values. As said in the simulation protocol, those statistics are mean, standard deviation and coefficient of a linear regression applied on data.

In summary for the results, if we consider mean as parameter and all the percentage of missing values, the methods estimating mean with a small bias are Listwise deletion, Mean imputation and Deterministic Regression imputation. The worst is k-NN with the biggest bias no matter the percentage of missing values observed.

Considering the standard deviation as parameter, again here Listwise deletion and Mean imputation are among the best methods in estimation. Close to them, Multiple Imputation can be added as good imputation method to estimate standard deviation. The worst method here is regression imputation without randomness with the biggest bias considering all the percentage of missing values.

For the coefficient of linear regression, the Listwise deletion and Mean imputation are on top of the methods,followed closely by regression imputation. Looking at the worst imputation methods used to estimate the coefficients, k-NN is on top of the list.

3.3 General comments and discussions

The results presented here are results from simulated data using specific distributions, sample size of 1000 and 1000 replications which leads to quite interesting and good results especially with mean imputation and Listwise deletion. These results can change if these parameters are changed. For example, with a bigger sample size or a smaller sample size, the results can change. With 200 as sample size you cannot impute 60% with a risk of changing the nature of initial distribution while with a larger sample size (2000 for example) you can go up to 70 percent if you want depending on the method. This simulation shows that up to 60% of data missing, results are almost the same. Bias isalmost the same for all the percentages meaning that it is possible, in certain cases, to impute more than 50 percent of the data when they are missing.

In this work, we found that for imputation methods like regression, the better the $R^2$the betterwill be the imputation results. It is not good to use regression imputation when the covariates explain a few percentage of the dependent variable presenting missing data. Consequence will be a very bad reconstruction of data leading of course to bias in all other estimators.

For some cases, methods like mean imputation can be improved by conditional mean imputation. In case the variable to impute is quite link or determine by another variable, conditional mean imputation on that other variable is advised. It is the same case for k-NN imputation which in this study did not perform very well because all the variables were generated randomly without link which is rarely the case in the true data sets.

As we have seen also in this simulation study, the sample size is quite big and we went up to 1000 replication to make sure of the stability of results. With a real data set, the statistician should rely on bootstrap to soften the bias that may occur during imputation. In addition, he/she should go for imputation methods that allows randomness like random regression imputation and multiple imputation.

The main conclusion or output drawn from the simulation section is a process to identify which method is suitable for imputation given a dataset. The process is as follows: use the variable in your dataset with missing data that you want to impute, truncate your data set and use only available cases to run the previous simulation process. This means that in the full matrix of the truncated data set, create missing values in the variable of interest and impute them using different methods. The method that gives you the best results will be used in the initial dataset to impute the values that are really missing. The algorithm to perform the best imputation with a real dataset is as follows:

- **Step 1**: Identify which variable in your dataset (Y) you would like to use imputation on, compute the percentage of missing values (s%) and identify all other variables that are determinant to Y in your data set.
- **Step 2**: Truncate your initial dataset and consider only case with all observations, a kind of complete case analysis. In this secondary data set, perform the simulation explained early in this section with s% of missing data. In other words, in the secondary data set without missing data, create s% of missing data in Y and impute them and compute RMSE and

MAE, perform it 1000 times to get standard deviation. The best method is the one that gives smallest values of RMSE and MAE.

- **Step 3**: Using the best method identified in step 2, perform imputation once in the initial dataset of step 1.

The results obtained from this process are surely the best we can get for imputation.

4. Applications

After simulations, the output of the analysis is a process to identify which method to use when we have a real data set. This section presents an application of this process.The data used here are from an agricultural household survey in Rwanda on 406 farming household over 4 regions in Rwanda. The variables of interest here is the Production of beans in Kg during wintering season of the year 2016 – 2017. Among the covariates we have: Use of climateinformation, Quantity of labour used, Quantity of seeds, Area cultivated, Tropical Livestock Index and Asset index. We applied the process described at the end of the section 4 and the results are summarized below[1].

4.1 Reconstruction of data

As in the simulation section, the reconstruction of data is measured by RMSE and MAE parameters. Figure 3 presents the change in RMSE according to each imputation methods and an increasing percentage of missing values.

[1] See appendices section for full results

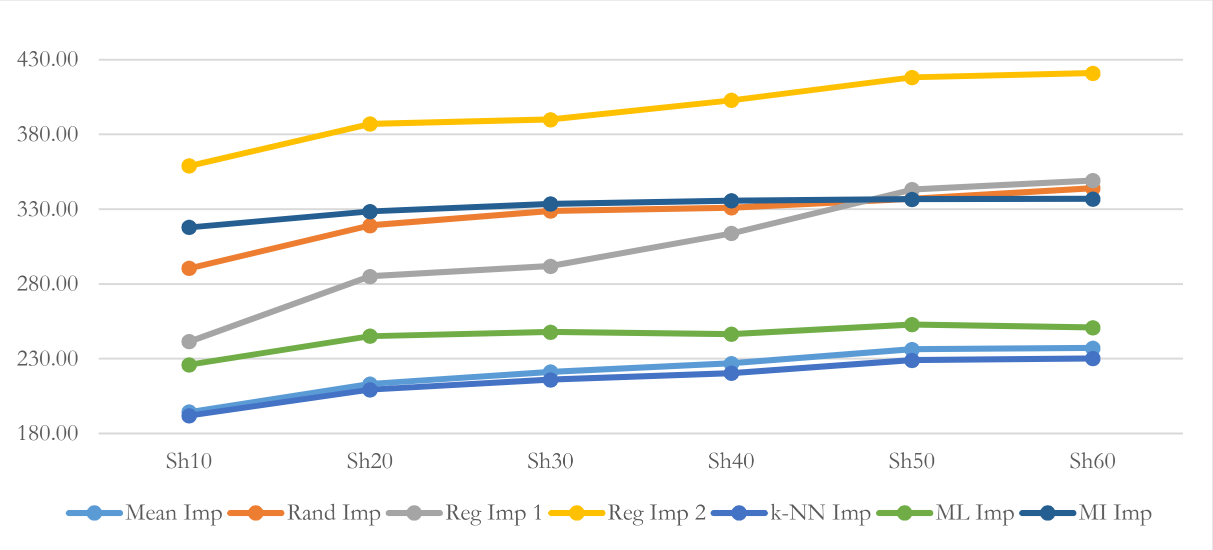

Figure 3: RMSE for imputation methods on real data of Rwanda

It is clear that for all the imputation methods, the RMSE increase with the percentage of missing values and the best method in reconstructing data is k-NN for this given data set. The second best is Mean imputation and the worst method is regression imputation with randomness.

If we look at the second indicator of goodness-of-fit in reconstruction in figure 4, the MAE is quite stable with the increasing percentage of missing value and it decrease even for Regression imputation and Multiple Imputation.

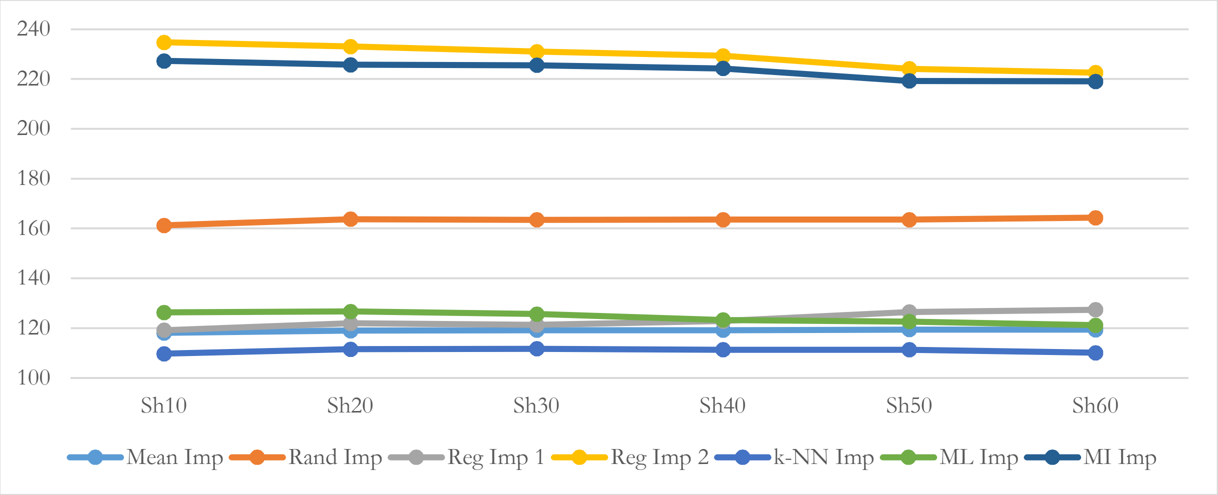

Figure 4: MAE for imputation methods on real data of Rwanda

Here again the best method in data reconstruction is k-NN and the second best is mean imputation.

The worst methods are Random regression imputation and Multiple imputation.

The conclusion here is that for this given dataset, in case imputations have to be made to reconstruct the data, the suitable methods are k-NN applied very well and Mean imputation.

4.2 Estimators of some statistics

If we look at the statistics estimated by the complete case analysis and the imputed dataset for each method on our variable of interest,the following conclusions can be drawn:

- ✓ In estimating the mean with the smallest bias, Listwise deletion, mean imputation and MI imputation are the three best imputation methods;
- ✓ In estimating standard deviation with the smallest bias, Listwise deletion, Multiple imputation and Regression imputation are the best three methods;
- ✓ In estimating coefficients of the linear regression, Regression imputation deterministic and random are the best methods.

Depending on what exactly you want to generate with your data, some methods are better than others. In absolute necessity of imputation, Multiple imputation will be the best one in estimating specific statistics with this dataset.

5. Conclusion

The aim of this study was to analyse the performance of imputation methods in case of simulated data and in case of real data. Finally, the main result obtained is that the performance of Imputation methods is closely link to the parameters of simulation and to the structure of data. Thus, an absolute decision cannot be taken. A major result here is that using bootstrap, the percentage of missing data in the variable doesn't matter much. We imputed up to 60% of missing data with quite good results in this study both in simulated and real dataset.

Practically, this study is more about explaining the process required to calibrate and identify which method will give better results during imputations in case data are missing completely at random. It cannot be used to compare imputation methods and conclude. In fact, as we have seen in simulations and applications, the methods performing very well are different depending on the simulation parameters and on the structure of the data when data are simulated. Even in case of real data, performance can change according to the profile of data (what are the different distributions concerned? are we having extreme values? Atypical values?). This study shows essentially in a case of missing data in a dataset, how to calibrate and choose which method will give you the best results.

More example of simulation and data set can be done using the simulation protocol developed here. There are many other imputation methods that can be tested. Given that bootstrap is used and 60% of data can be estimated using the methods tested in this work, imputation methods can be used beyond simple missing data estimation but also for censored data to estimate counterfactual in the framework of impact evaluation.

References

Afifi, A.A. &R.M. Elashoff (1966). Missing observations in multivariate statistics I. Review of the literature. Amer. Statistical Association 61: 595-604.

Allison, P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods & Research, 28(3), 301–309.

Allison, P. D. (2012). Handling Missing Data by Maximum Likelihood. SAS Global Forum 2012 Statistics and Data Analysis, 1–21.

Allison, P. D. (2003). Missing Data Techniques for Structural Equation Modeling. Journal of Abnormal Psychology, 112(4), 545-557.

Allison, P.D. (2001). Missing Data, Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-136, Thousand Oaks.

Briggs, A., Clark, T., Wolstenholme, J. &Clarke, P. (2003). Missing.... presumed at random: cost-analysis of incomplete data. Health Econ., 12: 377–392.

Gelman, A., & Hill, J. (2006). Missing-data imputation. In Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research, pp. 529-544). Cambridge: Cambridge University Press.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, Biometrics**27**, 857–874.

Hartley, H.O. and R.R. Hocking (1971). The analysis of incomplete data. Biometrics, 27: 783-823.

Kowarik A., & Templ M. (2016). Imputation with R package VIM. Journal of Statistical Software, 74(7): 1-16.

Little, R.J.A. & Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.

Little, R.J.A. & Rubin, D.B. (2002).Statistical Analysis with Missing Data, New York.

Orchard, T. & Woodbury M. A. (1972). A missing information principle: theory and applications. Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability, Theory of Statistics, Univ. of California Press.

Rubin, D.B. (1987).Multiple Imputation for Nonresponse in Surveys. New York, Chichester.

Rubin, D.B. (1996). Multiple Imputation after 18+ Years.Journal of the American Statistical Association, 91, 434, 473-489.

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

Schmitt, P.; Mandel, J. and Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. J Biomet Biostat 6: 224. doi:10.4172/2155-6180.1000224

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values.

Soley-Bori, M. (2013). Dealing with missing data: key assumptions and methods for applied analysis.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. and Hastie, T. (2001). Missing value estimation methods for dna microarrays. Bioinformatics 17: 520-525.Lewis HD.