Review on General Techniques and Packages for Data Imputation in R on a Real World Dataset

Fitore Muharemi[1(✉)], Doina Logofătu[1], and Florin Leon[2]

[1] Frankfurt University of Applied Sciences, Frankfurt Am Main, Germany
muharemi@stud.fra-uas.de, logofatu@fb2.fra-uas.de
[2] Technical University of Iaşi, Iaşi, Romania
florin.leon@tuiasi.ro

Abstract. When we collect data, usually they consist of small samples with missing values. As a consequence of this flaw, the data analysis becomes less effective. Almost all algorithms for statistical data analysis need a complete data set. In data preprocessing, we have to deal with missing values. Some well-known methods for filling missing values are: Mean, K-nearest neighbours (kNN), fuzzy K-means (FKM), etc. There are quite a lot of R packages offering the imputation of missing values, but sometimes its hard to find the appropriate algorithm for a particular dataset. When we have to deal with large datasets sometimes, these known methods cannot work as supposed because they need too much memory to perform their operations. This paper provides an overview of a considerable dataset imputation by applying three different algorithms. A comparison was performed using three different algorithms under a missing completely at random (MCAR) assumption, and based on the evaluation criteria: Root mean squared error (RMSE). The experiment results show that Random Forest algorithm can be quite useful for missing values imputation.

Keywords: Imputation · Missing values · Real data

1 Introduction

Missing data are a common problem in most scientific research domains such as Biology, Medicine, or Climatic Science. They can arise from different sources such as mishandling of samples, measurement error, non-response, or deleted value. Missing data based on three missingness mechanisms: data are missing completely at random (MCAR) when the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data; data are missing at random (MAR) when the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself; data are missing not at random (MNAR) when the probability of an instance having a missing value

N. T. Nguyen et al. (Eds.): ICCCI 2018, LNAI 11056, pp. 386–395, 2018.
https://doi.org/10.1007/978-3-319-98446-9_36

for a variable could depend on the value of that variable [9]. Missing data can affect statistical estimators such as means, variances or percentages, resulting in a loss of power and misleading conclusions. A variety of techniques have been proposed for substituting missing values with statistical prediction, and this process is generally referred to as 'missing data imputation' [8]. The dataset used for the experiment is a time series data, but here we completely ignored this fact, and in the present study, we compare three different imputation methods: K-nearest neighbours (kNN), multiple imputations by chained equations (MICE) and Random Forest. To evaluate the performance of the imputation techniques applied to this case study data, the Root Mean Square Error (RMSE) is used.

$$RMSE(\mathbf{a}, \mathbf{b}) = \sqrt{\frac{\sum_{i=1}^{n} (a_i - b_i)^2}{n}}$$

RMSE computes the distance between the imputed values and the originals - for categorical attributes, the distance is considered 1. Lower RMSE values represent the better predictive accuracy of the imputation method.

2 Data Characteristics

The Heating System dataset was introduced in GECCO IC 2015, and it is a real-world Data: Multiple real minutely heating system operating time series which are provided for training, testing and assessing data recovery methods[1]. Heating Systems are composed of a boiler which heats the water in the system, a pump to circulate the water and radiators which are wall-mounted panels through which the heated water passes in order to release heat into rooms. The circulating water systems use a closed loop, and the same water is heated, pumped through the heating circuit and then reheated again. Today modern heating systems log the full details of data to enable further interpretation of the data [3].

The data for the GECCO 2015 Industrial Challenge contains four different time series denoting the main aspects of the heating system behaviour. Given is the heating water temperature when leaving the boiler as well as the return temperature. Additionally, the setpoint for the heating water is supplied as well as the actual required power to heat up the water. The heating system that provided the data only logged the data when a value changed. Additionally, the data has been revised to supply periodic, minutely data. The heating system dataset is considered a large dataset as it contains 606837 data and 5-time series variables. A first impression related to these time series is given in Fig. 1. A seasonal pattern is observed in this time series. A seasonal pattern exists when a series is influenced by seasonal factors like the quarter of the year, the month, or day of the week. Seasonality is always of a fixed and known period. Hence, seasonal time series are sometimes called periodic time series.

[1] http://www.spotseven.de/gecco/gecco-challenge/gecco-challenge-2015/.

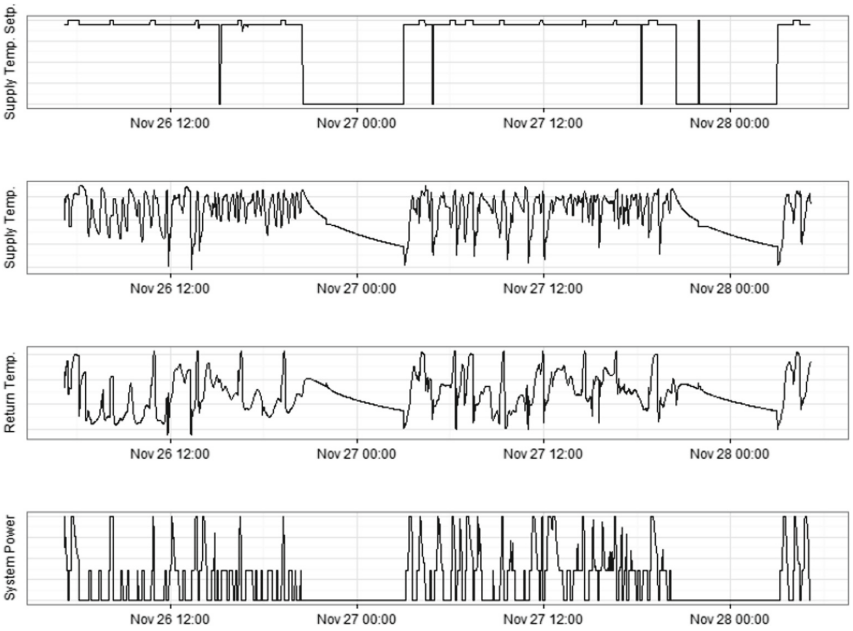

Fig. 1. Two days of the given time series data. The first row displays the target temperature for the water when leaving the boiler, while the second row shows the temperature the water has when leaving the boiler. The third row displays the temperature of the water when returning to the boiler, and the last row shows the amount of energy, spent that moment, to reheat the water.

3 Imputation Methods

Replacing the missing values with the mean, median or mode is a crude way of treating missing values. Depending on the context, like if the variation is low or if the variable has little leverage over the response, such a rough approximation is acceptable and could give satisfactory results.

kNN is an algorithm that is useful for matching a point with its closest k neighbours in a multi-dimensional space. kNN imputation is an imputation technique based on kNN algorithm designed to find k nearest neighbors for a missing instance of data (incomplete example) from all cases complete in a given dataset, and then fill in the missing values of example with the most frequent one occurring in the neighbors if the target feature (or attribute) is categorical, referred to as majority rule, or with the mean of the neighbors if the target feature is numerical, referred to mean rule. The kNN imputation approach has successfully been used in real data processing applications, kNN imputation is a lazy and instance-based estimation method and is one of the most common imputation techniques due to its simplicity, easy-understanding and relatively high accuracy. Different from model-based algorithms where the estimation is split into two phases: training phase where the estimator is build using the

complete dataset and then prediction phase where estimator is used to predicting missing values, kNN lacks the training phase and every estimation needs to be done using the complete dataset. While kNN imputation with majority/mean rule is simple and effective in general, there are still many efforts focusing on improving its performance. It can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data. The assumption behind using kNN for missing values is that a point value can be approximated by the values of the points that are closest to it, based on other variables. With kNN we need to consider some parameters like the number of neighbours - k, a low k will increase the influence of noise and the results are going to be less generalizable(overfitting). On the other hand, taking a high k will tend to blur local effects which are precisely what we are looking for. Another parameter is also the distance function: mostly based on Minkowski distance or Euclidian distance, but often these distance functions do not perform well for categorical variables while being productive on numerical ones. How kNN imputation works in practice is that for every observation to be imputed, it identifies 'k' closest observations based on the Euclidean distance and computes the weighted average (weighted based on distance) of these 'k' observations [5].

Data removed in	Interval		
No data removed	2013-11-18 05:12:00	–	2014-01-24 11:02:00
Supply temperature setpoint	2014-01-24 11:03:00	–	2014-03-24 11:43:00
System supply temperature	2014-03-24 11:44:00	–	2014-05-22 13:24:00
Return temperature	2014-05-22 13:25:00	–	2014-07-20 14:05:00
System power	2014-07-20 14:06:00	–	2014-09-17 14:46:00
All 4 time series	2014-09-17 14:47:00	–	2014-11-15 14:27:00
No data removed	2014-11-15 14:28:00	–	2015-01-13 15:08:00

Fig. 2. Based on the missing values, data can be divided into 7 intervals, one leading and one trailing interval with complete data, and five intervals showing gaps.

Multiple Imputation. There are two basic imputation strategies: single imputation which provides single estimation for each missing data and multiple imputations such as multiple imputations (MI) (Little and Rubin 2002) and EM algorithm (Dempster et al. 1977) [1,4]. Some popular single imputation methods include hot deck imputation and mean imputation. Hot-deck imputation replaces missing values with responses from other records that satisfy certain matching conditions. Mean imputation estimates missing values by the mean value of appropriately selected "similar" samples. A limitation of single imputation strategy is that it tends to reduce the variability of characterizations of the imputed dataset artificially. With single-imputation techniques, missing values in a variable are imputed by an estimated value that results from matching some specific conditions. Single imputation cannot provide valid standard errors and confidence intervals since it ignores the implicit uncertainty. Furthermore,

imputing the missing value with a single value does not capture the sample variability of the imputed value or the uncertainty associated with the model used for imputation. The alternatives of single imputation are to fill in the missing values with multiple imputations or iterative imputation such as MI or EM algorithm. MI algorithm assumption requires that the data (explicitly missing data) should be of type MCAR (missing completely at random) to generate a general-purpose imputation. In multiple imputation strategies, several different imputed datasets are, and a set of statistical results can be computed from it. Strength of the Amelia package lies in allowing multiple imputations for time series data. It uses bootstrapping and Expectation-Maximization algorithm, to impute the missing values in a dataset. Dataset is copied as many times we want as shown below. Multiple imputations involve imputing m values for each missing cell in your data matrix and creating m "completed" data sets. If the number of imputations we specified is 3, then it will create three copies of the original dataset and then using EM algorithm, the missing values for each imputed set is calculated, so we end up with 3 complete datasets. Across these completed data sets, the observed values are the same, but the missing values are filled in with different imputations that reflect our uncertainty about the missing data. After imputation, Amelia will then save the m-datasets. You then apply whatever statistical method you would have used if there had been no missing values to each of the m-datasets, and use a simple procedure to combine the results It generalizes existing approaches by allowing for trends in time series across observations within a cross-sectional unit, as well as priors that allow experts to incorporate beliefs they have about the values of missing cells in their data. Amelia II also includes useful diagnostics of the fit of multiple imputation models. The advantage of Amelia is that it combines the comparative speed and ease-of-use of the algorithm with the power of multiple imputations, to let us focus on your substantive research questions rather than spending time developing complex application-specific models for nonresponse in each new data set. Unless the rate of missingness is exceptionally high, $m = 5$ (the program default) will usually be adequate. When multiple imputations work properly, it fills in data in such a way as to not change any relationships in the data but which enables the inclusion of all the observed data in the partially missing rows. Amelia II implements the bootstrapping-based algorithm that gives virtually the same answers as the standard MI or EM approaches, is usually considerably faster than other approaches. The assumptions Amelia makes are: dataset is multivariate normal and missing data values belong to MAR(Missing At Random). By assuming that time series data vary smoothly over time, observed values close in time to the missing value can significantly aid imputation of that value. However, the trend pattern may change over time.

Random forest (RF) missing data algorithms are an attractive approach for dealing with missing data because of the desirable properties of being able to handle mixed types of missing data, they are adaptive to interactions and nonlinearity, and they have the potential to scale to big data settings. Studies reveal that RF imputation to be generally robust with performance improving

with increasing correlation, good performance in moderate to high missingness, and in some cases even when data was missing not at random. By averaging over many unpruned classification or regression trees, random forest intrinsically constitutes multiple imputation schemes. Using the built-in out-of-bag error estimates of random forest, we can estimate the imputation error without the need for a test set. Related studies show that missForest can successfully handle missing values, particularly in datasets including different types of variables. Furthermore, sometimes missForest outperforms other methods of imputation especially in data settings where complex interactions and non-linear relations are suspected. The out-of-bag imputation error estimates of missForest prove to be adequate in all settings. Additionally, missForest exhibits attractive computational efficiency and can cope with high-dimensional data.

Random forests replaces missing values only in the training set, and it begins by doing a rough and inaccurate filling in of the missing values. After this filling it does a forest run and computes proximities. If *x(a,b)* is a missing continuous value, estimate its fill as an average over the non-missing values of the *a-th* variables weighted by the proximities between the *b-th* case and the non-missing value case. If it is a missing categorical variable, replace it by the most frequent non-missing value where frequency is weighted by proximity. After that iterate-construct a forest again using these newly filled in values, find new fills and iterate again. Normally 5–6 iterations are considered enough [2].

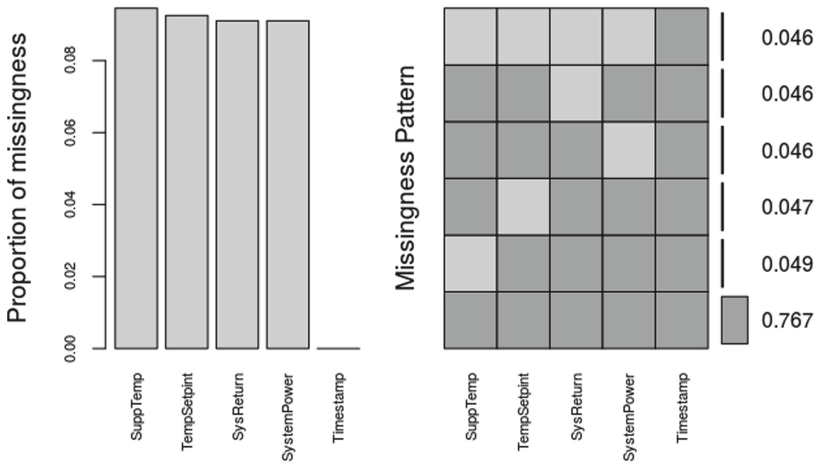

Fig. 3. Visual representation of missingness pattern

The advantage of the RF is that it is very robust non-parametric multiple imputation technique that can handle any input data without making assumptions about structural aspect of the data itself. We choose RF because it performs very well with mixed types of data under barren conditions like high dimensions, complex interactions and non-linear data structures. The missing data problem

is addressed using an iterative imputation scheme by training an RF on observed values in a first step, followed by predicting the missing values and then proceeding iteratively. Due to its accuracy and robustness, RF is well suited for the use in applied research often harbouring such conditions. Furthermore, as we mentioned the RF algorithm allows for estimating out-of-bag (OOB) error rates without the need for a test set [7].

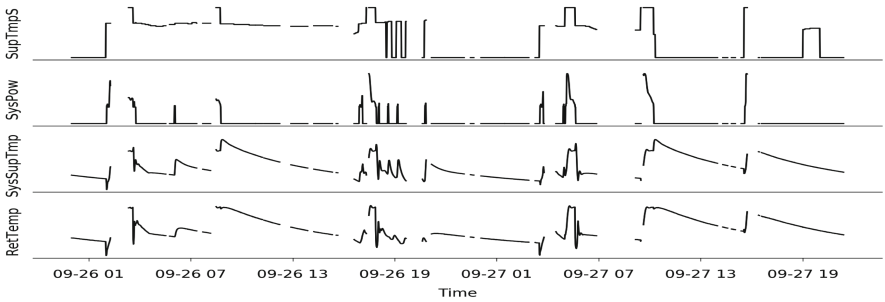

Fig. 4. Two days data with missing values

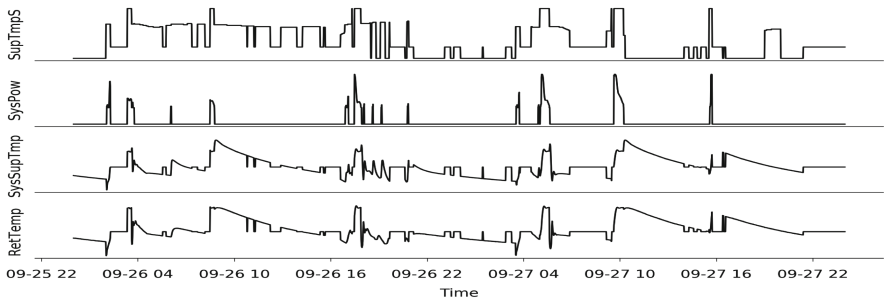

Fig. 5. Two days data after imputation with random forest algorithm

4 Analysis

Figure 2 introduces the dataset with missing values. It is beneficial knowing and having this kind of information. In this paper, we assume that we don't know any information and we try what methods can achieve the best result. In Fig. 3 is visualized that there are 76% values in the data set with no missing value. There are 4.9% missing values in Supply_Temperature, 4.7% missing values in Temperature_Setpoint and so on. We can also look at the histogram which depicts the influence of missing values in the variables.

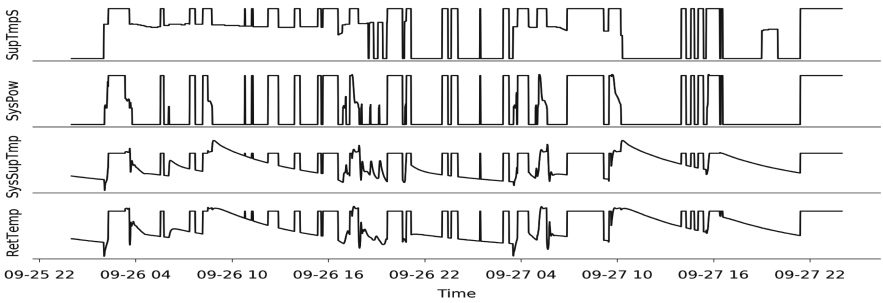

Fig. 6. Two days data after imputation with kNN algorithm

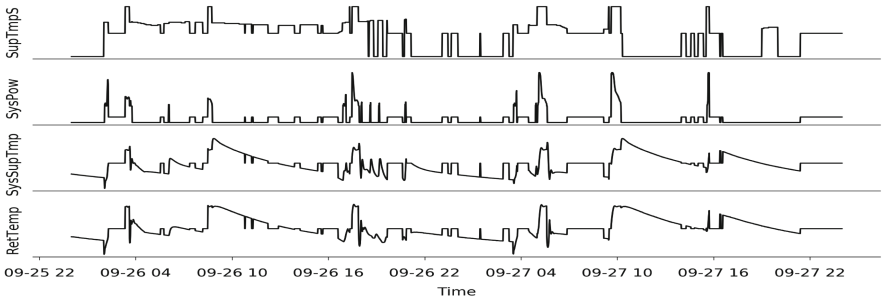

Fig. 7. Two days data after imputation with multiple imputation

5 Evaluation Criteria

The root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values observed.

It indicates the absolute fit of the model to the data - how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of the variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit than the larger ones. RMSE is a good measure of how accurately the model predicts the response and is the most crucial criterion for fit if the primary purpose of the model is the prediction. Mean Absolute Error (MAE), and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for continuous variables. Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to infinite and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to significant errors. The RMSE should be more useful when large errors are particularly undesirable, so RMSE has

the benefit of penalising large errors more so can be more appropriate in some cases. One distinct advantage of RMSE over MAE is that RMSE avoids the use of taking the absolute value, which is undesirable in many mathematical calculations that we will not discuss here [6].

6 Results

Three different imputation methods were selected to compare them by imputing the missing values in a sizeable real-life dataset, heating system data. As we ignored many pieces of information we had on this dataset to try Random Forest, Multiple Imputation and kNN algorithms, it was not predictive what algorithm can give most accurate results. Figure 4 plots a two-day interval of this dataset followed by Figs. 5, 6 and 7 which show the imputed dataset with three different algorithms.

According to RMSE criteria, Random Forest algorithm outperforms other two algorithms by giving an RMSE = 5.1, while in MI we achieved the RMSE = 6.2. **kNN** algorithm (three different k = 3, 5, 7) was less effective method when applied to the heating system dataset (Table 1).

Table 1. RMSE of imputed dataset with three different algorithms

Algorithm	kNN	Random forest	Multiple imputation
RMSE	13.1	5.1	6.2

7 Conclusion and Future Work

Missing data are almost a crucial part of data science research. There are several alternative ways to deal with missing values and no general algorithm performs well on all different datasets. However, only a few studies report an evaluation of existing imputation methods. In the present study, we performed a neutral comparison of three imputation methods based on one large real dataset, under an MCAR assumption. Validation of imputation results is an important step and most often for the evaluation is considered the criteria: Root mean squared error (RMSE). While much attention has been paid to the imputation accuracy measured by RMSE, only a few studies have examined the effect of imputation on high-level analyses [8,11]. Better results can be achieved by using different techniques on this dataset but we intend to provide a general conclusion independent from the domain of application, and we could certainly further improve the accuracy of imputation by integrating specific imputation methods.

Further development of other techniques would be helpful in the field of data science, as the imputation is the first step to face during data analysis. This study has many limitations, but as future works, we try to find a model that

highlights some attributes and decreases the impact of the rest on the missing datum. Find specific completed instanced to use for kNN imputation and not all the dataset. Recently, Zhang (2008, 2011) proposed a new imputation direction, named parimputation strategy, in which a missing datum is imputed if and only if there are specific complete instances in a small neighbourhood of the missing datum [12].

References

1. Allison, P.D.: Missing data: quantitative applications in the social sciences. Br. J. Math. Stat. Psychol. **55**(1), 193–196 (2002)
2. Breiman, L.: Random forests Leo Breiman and Adele Cutler. Random Forests-Classification Description (2015)
3. Christopher, F., Thomas: Gecco 2015 recovering missing information in heating system recovering missing information in heating system operating dataoperating data (2015)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**, 1–38 (1977)
5. Faisal, S., Tutz, G.: Nearest neighbor imputation for categorical data by weighting of attributes. arXiv preprint arXiv:1710.01011 (2017)
6. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M.: Methods for imputation of missing values in air quality data sets. Atmos. Environ. **38**(18), 2895–2907 (2004)
7. Mitchell, M.W.: Bias of the random forest out-of-bag (OOB) error for certain input parameters (2011)
8. Schmitt, P., Mandel, J., Guedj, M.: A comparison of six methods for missing data imputation. J. Biometrics Biostatistics **6**(1), 1 (2015)
9. Shrive, F.M., Stuart, H., Quan, H., Ghali, W.A.: Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Med. Res. Methodol. **6**(1), 57 (2006)
10. Troyanskaya, O., et al.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)
11. Wang, D., et al.: Effects of replacing the unreliable cdna microarray measurements on the disease classification based on gene expression profiles and functional modules. Bioinformatics **22**(23), 2883–2889 (2006)
12. Zhang, S.: Nearest neighbor selection for iteratively kNN imputation. J. Syst. Softw. **85**(11), 2541–2552 (2012)