Proceeding of IC-ITS 2018 e-ISBN:978-967-2122-64-7
4th International Conference on Information Technology & Society
October 30th-31st, 2018, Selangor, MALAYSIA

Dealing with Missing Data using Two R Packages in RGB Colour Histogram Data

Juzlinda Mohd Ghazali and Aisyah Mat Jasin

International Islamic University College Selangor (KUIS)

juzlinda@kuis.edu.my, aisyahjasin@kuis.edu.my

Abstract. Multiple imputation (MI) is a powerful tool in handling missing data issue. This paper provides a comparison of the multiple imputation method in Amelia II package and MICE package in R. Both packages are well-known and incredible to conduct the missing data research in numerous domains. There are very limited researches comparing the multiple imputation combined with other techniques in the image data context. We employ the mean absolute error (MAE) error metric to evaluate the accuracy on the predicted values based on 20% and 50% of missing data percentages. Although the implementation of MICE is time consuming, the result shows that MICE can deal with large amount of missing values while Amelia II is only capable to deal up 20% amount of missing values. In the MAE result, both packages show that they are superior on the particular variables.

INTRODUCTION

Missing data is the common problem in the data quality issues. It will give a significant impact to the statistical inference on data analysis result. Therefore, the missing data problem have been received a great concern from professional researchers to address this issue that associated to decision making and planning.

LITERATURE REVIEW

Numerous efforts have been introduced to improve the existing solutions of the missing data problem [1]. Traditionally, deletion method is one of the easiest technique in handling missing data solution. Later, imputation method was introduced where the missing values are substituted with plausible values for instance; mean, median and mode values. Although the mean imputation technique is may enhance drawbacks in deletion technique, the results of the imputed data are obviously bias. The conditional mean and stochastic the invented for improving bias. All these aforementioned techniques are categorized of the single imputation methods [2].

The multiple imputation [3] is one of the strategies to overcome number of weaknesses in the single imputation technique. It was first introduced by Rubin in 1978 and the early developments of the multiple imputation method. Multiple imputation is an iterative procedure and consist three distinct phases: (i) Imputation phase creates several copies of imputed data sets (says, m=5) where missing values are imputed by plausible values commonly applied iterative stochastic regression imputation. Each of m copy of data set will contain different imputed missing values. (ii) Analysis phase estimates the statistical inference such as parameter estimates $\theta = \mu, \Sigma$ and standard errors of each imputed data. Therefore, each m of imputed data yield different m as parameter estimates and standard errors. (iii) pooling phase combines m as parameter estimates and standard errors into a single parameter estimate and standard error [4].

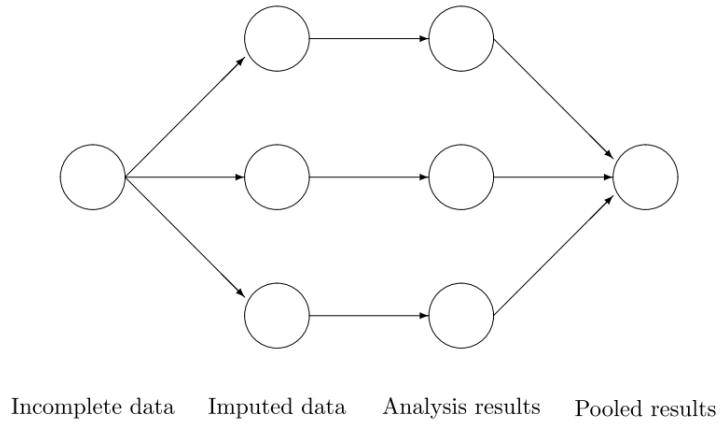


FIGURE 1. Scheme of Main Steps in Multiple Imputation [5].

The purpose of this paper is to provide a simulation study of the multiple imputation approach on the image data. We differentiate the performance of the two well-known multiple imputation packages in R called Amelia II [6] and Multiple imputation by Chain Equation (MICE) [7] in which each package employed different statistical approaches. The first package is called Amelia II combines the expectation maximization bootstrap (EMB) algorithm; it is able to impute high dimensional missing data with less time. This package implemented by Honaker and King in 2010. The second package is MICE which is also known as fully conditional specification or sequential regression multiple imputation. This package provides several extension procedures that combined with multiple imputation process. However, authors specified the predictive mean matching approach as an extension approach of the multiple imputation technique employed in the MICE package and we applied it in this simulation study.

METHODOLOGY

Let X be the n x p matrix data where n is the number of observations and p is the total number of variables or components. We assumed X is a matrix of multivariate distribution that completely specified by unknown parameter θ. We denote the X_{obs} and X_{miss} are observed components and missing components respectively. The standard multiple imputation algorithm in Amelia II and MICE are as follow [8]:

1. Estimate θ from the posterior distribution $p(\theta|\mathrm{X}_{obs})$ based on the observed data y_{obs}
2. Estimate θ from $p(\theta|\mathrm{X}_{obs})$
3. Draw a value x of X_{miss} from the conditional posterior distribution $p(X_{miss}|\mathrm{X}_{obs}, \theta)$ given by X_{obs}
4. and θ

The parameters estimation θ are obtained by sampling iteratively by conditional distribution $p(X_1|\mathrm{X}_{-1}, \theta_1) \ldots p(X_p|\mathrm{X}_{-p}, \theta_p)$.

In this simulation, the missing data are assumed to missing completely at random (MCAR) where the missingness are not rely on any variables in the data set. The missing data pattern presented in the Figure 2.

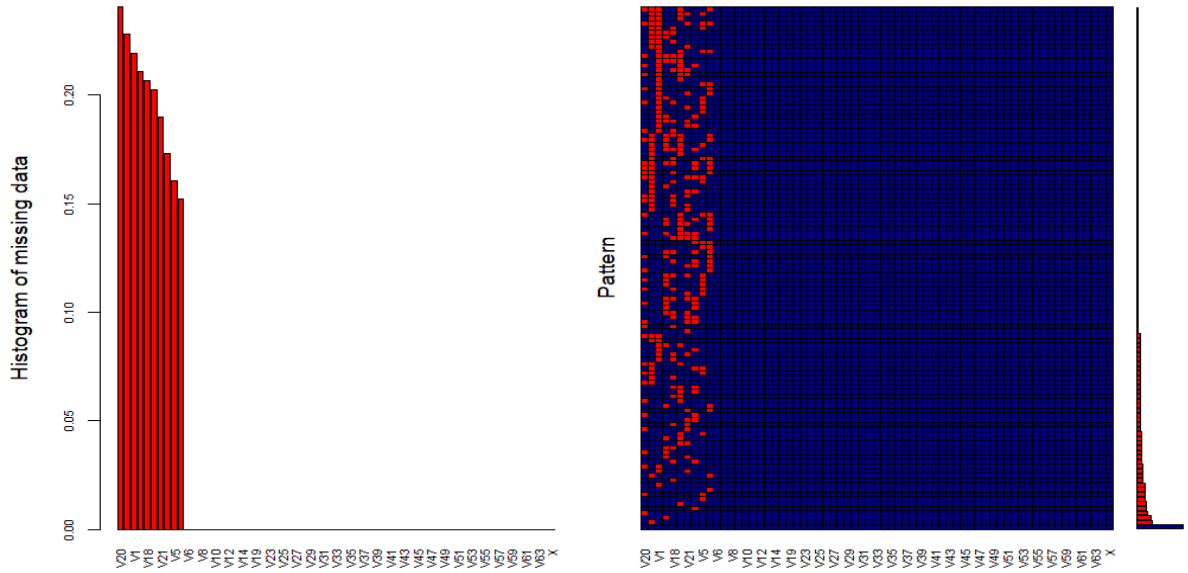


FIGURE 2. Missing Data Pattern or Missingness Map of the 64-Variables of the RGB Colour in the Image. The Red Spots Indicate that Missing Values Occur in The Selected Variables.

EXPERIMENTAL DESIGN

We compare two incredible multiple imputation packages in R that able to work with continuous variables. With this simulation study, we may learn which package that suitable to deal with the high dimensional continuous data set.

We applied the multiple imputation technique using Amelia II and MICE package in R on a dataset contains RGB colour histogram whereby the colour feature extracted from 4 different images: (i) building, (ii) festival, (iii) beach, (iv) mountain. The colour histogram generated provides the distribution of RGB colours in the image. Generally, the 3D RGB colour space is divided into cells and for each cell, the number of pixels is counted.

A 3D RGB colour space is projected by each pixel in the image, as illustrated in Figure 1. The 3D colour space was divided into 4x4x4 cells which generate colour histogram with 64 bins. The number of pixels in each cell is counted and stored in the colour histogram. The total number of pixels in each bin is added up to get the total value. Each bin value is divided by the total value. This normalized data gives the proportion of pixels as a percentage for each bin.

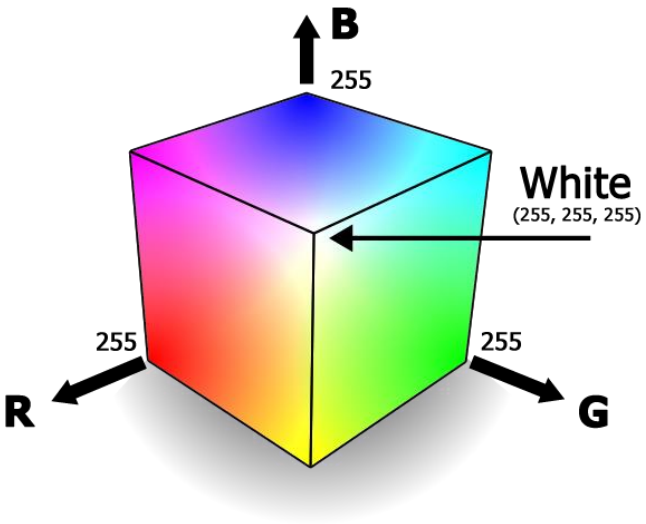


FIGURE 3. Illustration of RGB Colour Cube.

The purpose of this simulation study is to measure the performance and prediction accuracy between predicted and actual values. The mean absolute error (MAE) was used to measure the average error between predicted and actual values. The greater the deviation means the larger error between predicted and actual values and lower shows better result. The two evaluation criteria are:

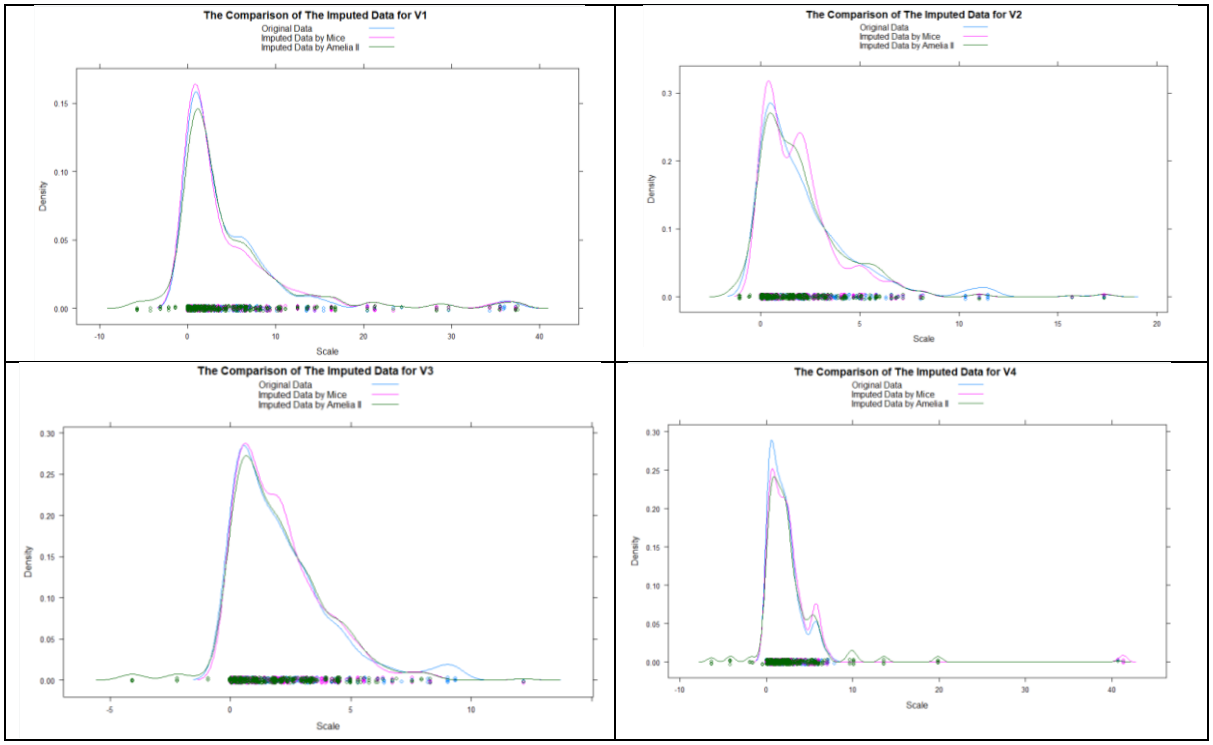
$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} \left| x_i - \hat{x}_i \right| \tag{13}$$

EXPERIMENTAL RESULT AND CONCLUSION

We applied the MAE error metric as an evaluation criterion to assess the accuracy of the imputed data. The result obtained in Table 1 shows that the multiple imputation method in Amelia R package outperformed the multiple imputation algorithm proposed in MICE. We assessed the performance of multiple imputation in both R packages based on 20% of missing data percentage. The result show that both Amelia II and MICE algorithm superior in the specific variables as shown with bold font. When we added the MDP from 30% up 50%, only MICE consistently work to impute the missing data. However, the Amelia is only works well if missing data percentage is less than 20%. Otherwise, there are some limitations appeared such as collinearity issues etc.

TABLE (1). The MAE And R-Square Estimates.

	MAE	
Variables	**Mice**	**Amelia**
VI	0.44791139	0.34072639
V2	0.31487342	0.21944868
V3	0.18050211	0.23295582
V4	0.52897890	0.40709509
V5	0.05290295	0.07450239
V16	0.03011392	0.03675646
V17	0.02764979	0.02926296
V18	0.01716456	0.02581327
V20	0.03094937	0.02281679
V21	0.02005907	0.01663878



REFERENCES

[1] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[2] A. Briggs, T. Clark, J. Wolstenholme, and P. Clarke, "Missing .... presumed at random : cost-analysis of incomplete data," *Health Econ.*, vol. 12, no. May, pp. 377–392, 2008.

[3] D. B. Rubin, "Multiple imputation in sample surveys - A phenomenological bayesian approach to nonresponse," in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1978, pp. 20–28.

[4] D. B. Rubin, "An overview of multiple imputation," in *Proceedings of the Section on Survey Research Methods*, 1988, pp. 79–84.

[5] S. Van Buuren, *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL, 2012.

[6] J. Honaker, G. King, and M. Blackwell, "AMELIA II : A Program for Missing Data," *J. Stat. Softw.*, vol. 45, pp. 1–47, 2011.

[7] S. van Buuren and K. Groothuis-Oudshoorn, "**mice** : Multivariate Imputation by Chained Equations in *R*," *J. Stat. Softw.*, vol. 45, no. 3, 2011.

[8] D. B. Rubin, "Fully conditional specification in multivariate imputation," *J. Stat. Comput. Simul.*, vol. 76, no. 12, pp. 1049–1064, 2006.