Comparative Statistical Algorithms for Imputation of Missing Measurements in Petrophysical Data

Conference Paper · December 2014

DOI: 10.13140/RG.2.1.2368.3285

2 authors:


Watheq J. Al-Mudhafar
University of Texas at Austin


Ali Al-Mudhafar
University of Basrah

Comparative Statistical Algorithms for Imputation of Missing Measurements in Petrophysical Data

W.J.M. Al-Mudhafar, Louisiana State University & A. Al-Mudhafar, Basrah University

The missing data problem in petrophysical properties especially core measurements of permeability is a crucial step in reservoir characterization. It affects the multivariate statistical inference of these measurements leading to non-efficient prediction and less accurate geospatial modeling.

Consequently, many imputation algorithms have been presented in this paper to comparatively predict the missing values of horizontal and vertical core permeability for a well in sandstone reservoir in a southern Iraqi oil field. The algorithms are Mean Substitution (MS), Iterative robust model-based imputation (IRMI), Multiple Imputation of Incomplete Multivariate Data (MIIMD), and Random Imputation of Missing Data (RIMD). These algorithms have been applied based on the deductive statistical inference to impute the incomplete data.

All the algorithms above have been illustrated and the predictions have been depicted for the data before and after the imputation process for all the algorithms with respect to the histograms and the vertical data distribution given the well depth. The results have shown that the Random Imputation of Missing Data is the best algorithm because the histogram has preserved its shape before and after the imputation. Therefore, it is the best one for accurate imputation of incomplete petrophysical data.

Introduction

The missing data in petrophysical properties is a common problem that should be efficiently handled to obtain accurate inference. The most conventional way is to consider the correlation between the porosity and permeability given the well log records; however, it might result in inaccurate prediction. Therefore, it is important to look for a modern algorithms to accurately impute the missing data in order to capture the real property distribution and to capture the reservoir heterogeneity in the spatial modelling.

In this paper, four algorithms have been adopted to deductively impute the missing measurements of horizontal and vertical core permeability for a well in sandstone reservoir, southern Iraqi oil field. The histogram and vertical distribution of data before and after the imputation have been done for comparison between the four algorithms.

Mean Substitution of Missing Data

Mean Substitution algorithm (Barladi and Enders, 2010) deals with the missing data by computing the mean for the related variable and it has the advantage of overcoming the probability of outlier points as shown in the figures below:

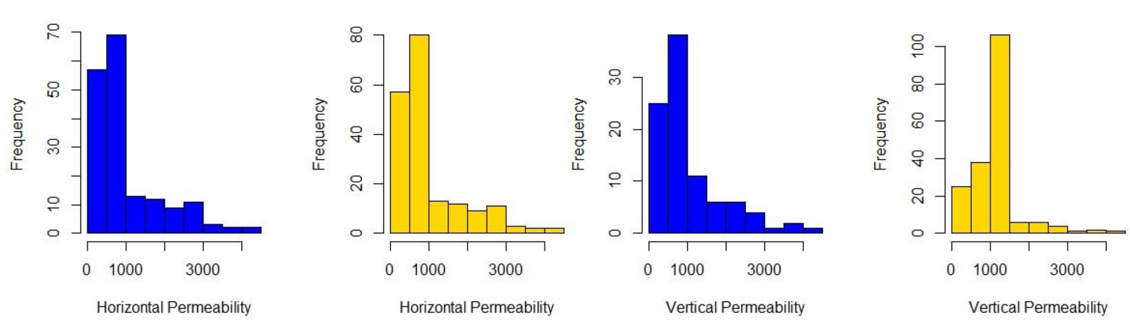


Figure 1: Comparison of histogram before (blue) and after (yellow) imputing the missing data by Mean Substitution Algorithm.

Iterative robust model-based imputation (IRMI)

This algorithm deals with missing data in iterative and sequential procedure of stepwise regression. The method can deal with a mixture of continuous and discrete variables (Temple r al., 2011).
In the given dataset and in each step of the iteration, one variable is considered as a response factor and other variables are treated as predictors to be fitted in the regression modeling. The missing data is imputed by obtaining the best fit equation. For the given data, the results have shown distinct distribution from its original as shown in the following histograms:-

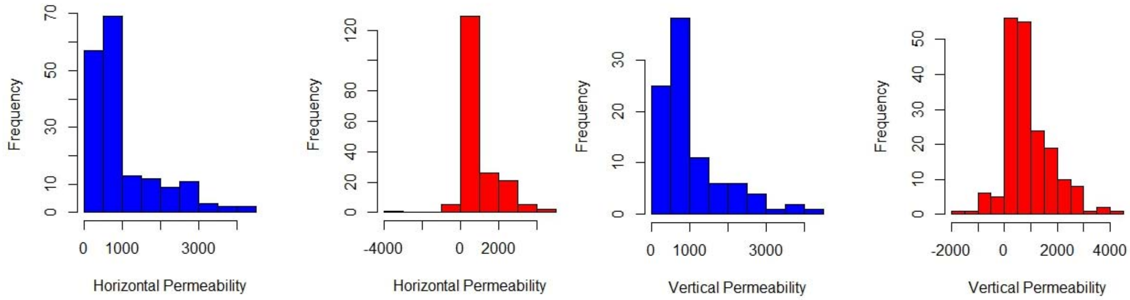
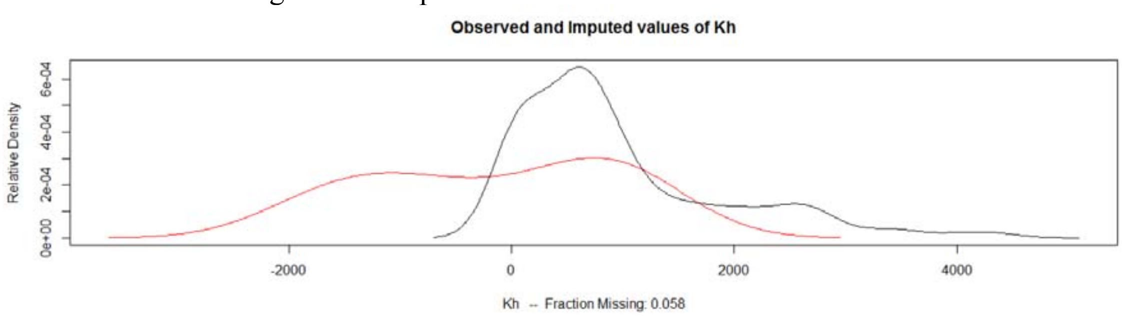


***Figure 2:** Comparison of histogram before (blue) and after (red) imputing the missing data by Iterative robust model-based imputation (IRMI) Algorithm.*

Multiple Imputation of Incomplete Multivariate Data

This algorithm deals with complete data parameters. It consider the Expectation Maximization (EM) algorithm with bootstrapping to compute the posterior probability distribution after estimate the likelihood function (Schafer and Olsen, 1998). Then, the imputed dataset is made by taking draws from the posterior mode for the complete dataset to the missing data given the available observation. The following figures show the data distribution for the original and imputed data.



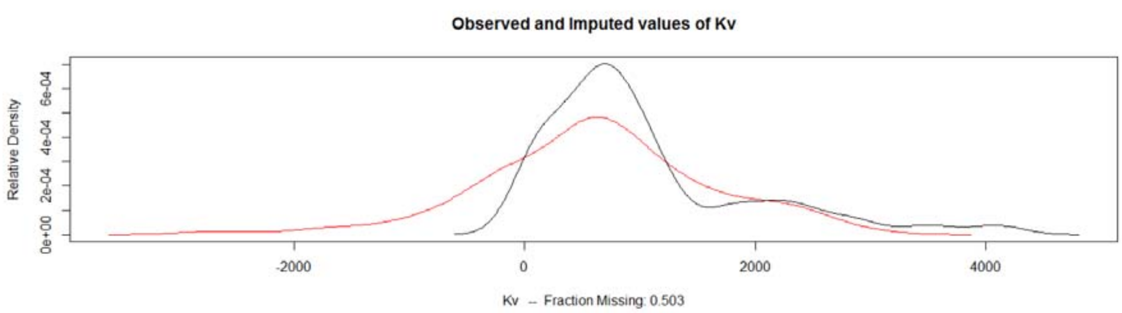


Figure 3: Comparison of density plots before (black) and after (red) imputing the missing data by Multiple Imputation of Incomplete Multivariate Data Algorithm.

Random Imputation of Missing Data (RIMD)

It is the simplest procedure for data imputation for only one variable based on the observed data for this variable. The procedure is repeated for every variable in the dataset (Gelman and Hill 2006). The imputed data that has been resulted from this algorithm has the most convenient outcomes compared to the other algorithms.

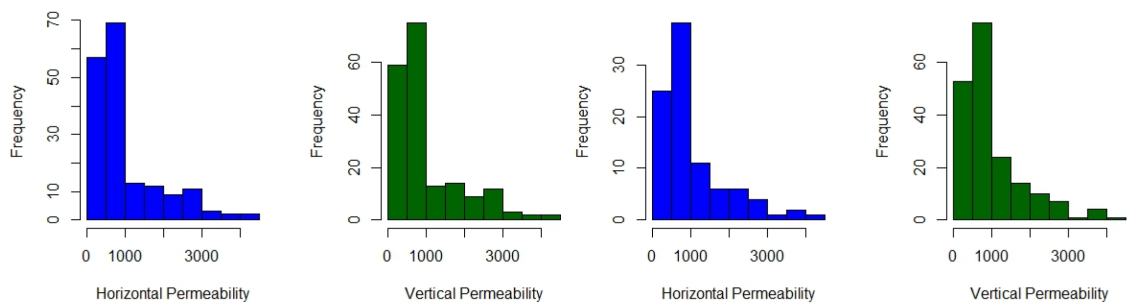


Figure 4: *Comparison of density plots before (blue) and after (green) imputing the missing data by Random Imputation of Missing Data (RIMD) Algorithm.*

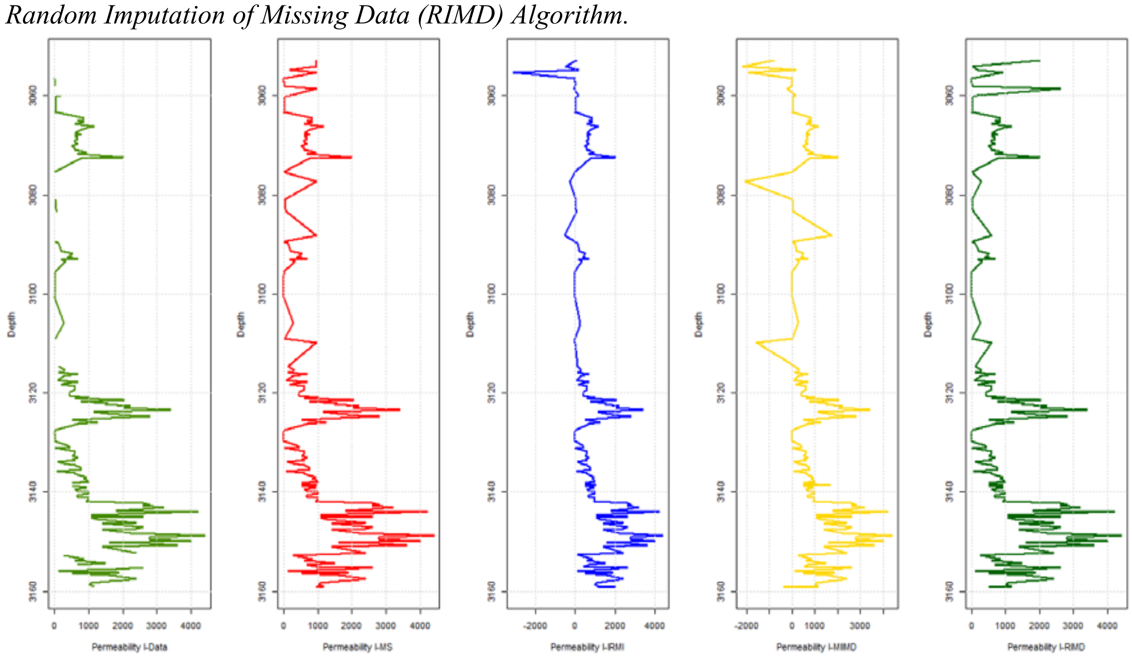


Figure 5: *Comparison of the horizontal permeability distribution between the original data (left) and the four imputation algorithms.*

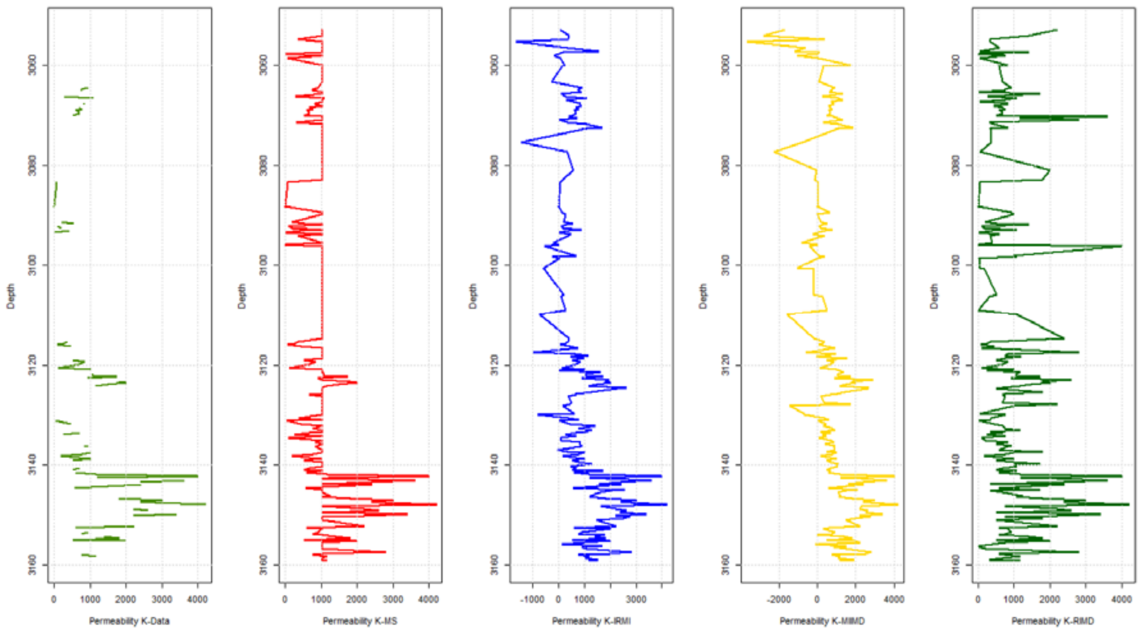


Figure 6: *Comparison of the Vertical permeability distribution between the original data (left) and the four imputation algorithms.*

Conclusions

From the diagnostic figures, it was prominent that the Random Imputation of Missing Data (RIMD) algorithm has the most accurate imputation of missing horizontal and vertical core permeability. That is clear from the histogram of the data before and after the imputation process. Also, the vertical data distribution by this algorithm is closer to the original data distribution than other algorithms.
Therefore, we can accordingly consider this algorithm to perform solid multivariate dataset to be more trustful basis for geospatial modelling.

References

1. Barladi, A. N. & C. K., Enders, (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48, 5-37*

2. Templ, M., A. Kowarik, and P. Filzmoser (2011) Iterative stepwise regression imputation using standard and robust methods. *Journal of Computational Statistics and Data Analysis*, Vol. 55, pp. 2793-2806.

3. Schafer, J. L., & M. K., Olsen, (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4), 545-571.

4. Rubin, D. B., & N., Schenker, (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), 366-374.

5. Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. *Cambridge University Press.*

6. Al-Mudhafer, W. (2013). Maximum Likelihood & Multiple Imputation of Incomplete Static and Dynamic Reservoir Data. *In Geoinformatics 2013.*