Gazi University Journal of Science
GU J Sci
24(1):69-75 (2011)

GUJS
www.gujs.org

A Comparison of Hot Deck Imputation and Substitution Methods in The Estimation of Missing Data

Abdullah YEŞİLOVA [1♠], Yılmaz KAYA [1], M. Nuri ALMALI [2]

[1] *Yuzuncu Yil University, Faculty of Agriculture, Biometry-Genetic Unit, Van, TURKEY*

[2] *Yuzuncu Yil University, Vocational High School, Computer Programming , Van, TURKEY*

[3] *Yuzuncu Yil University, Faculty of Engineering and Architecture, Department of Electrics and Electronics Engineering, Van, TURKEY*

Received: 19.03.2010 Revised: 13.09.2010 Accepted: 22.09.2010

ABSTRACT

It is of great importance to obtain data in an accurate and incomplete way for adequate conclusions to be drawn from investigations conducted. Due to various reasons, certain parts of an investigation might not be observed, and as a result of this, data might be missing and obtained incompletely. Missing value may not only be based on a single variable but also a multitude of variables. In this study, missing data in different proportions and belonging to more than a variable were produced. When data were considered within a context which is missing completely at random, Hot Deck imputation, random Hot Deck imputation and substitution methods (mean, median) were compared in the estimation of missing value. As a result of analysis, Hot Deck imputation method was found to be more effective in the estimation of missing value.

Key Words: *Missing data, Hot Deck imputation, Substitution methods.*

1. INTRODUCTION

Missing value in the analysis of statistical data is a frequently encountered problem. Missing value may belong to a single variable on a set of variables. In such cases, analyses are usually conducted in a way that missing data are excluded from the processing. Exclusion of missing data from the analysis may lead to biased parameter estimations [8]. In addition, the accuracy and generalization of the study are effected due to the missing data omitted from the data set [2,3].

Different methods have been devised for the estimation of missing value. Selecting an appropriate method depends on the mechanism playing a role in missing value. This mechanism has been divided into three categories. These are: missing completely at random (MCAR), missing at random (MAR) and non ignorable (NI) [7, 8]. In this study our analyses were conducted within the context of the method of missing completely at random when data are missed. In MCAR, if the missing value for X variable not dependent upon any variable or X variable, the missing value are taken within the context of MCAR. In the other words, from the data being based on missing observations are meant that such data are not dependent on any variable in the data set [7, 9]. Missing data in this set stem from reasons which may completely be coincidental.

♠Corresponding author, e-mail: yesilova@yyu.edu.tr

In this study, the software in C# programming language (shown in appendix) was devised for determining the performances of Hot Deck imputation, random Hot Deck imputation and substitution methods using the nearest k-neighbors algorithm in the estimation of missing value. In order to evaluate the effectiveness of methods, different correlation coefficients between the estimated data sets and real data sets were computed.

2. DATA SET

The data set used in this study is based on the principle of those used in the investigations conducted within the scope of the Project numbered as 2003-DPT-MIM1. In this study, microprocessor-controlled measurement devices were set up in Yüzüncü Yıl University campus in a height of 30 m and 10 m suited to the standards. The speed of the wind was measured in 30s period of time, the mean value and standard deviations of each 10 minute, as well as extreme values (max and min) were recorded. The data related with the speed of the wind were recorded in 10 minutes intervals and belong to one year of period of time between April -2004 and March 2005. In this study, variables belonging to January 2005 and listed in the following table were used.

Table 1. Variables in data set.

Code	Variable
S1	Wind speed in 30 meters of height
S2	Wind speed in 10 meters of height
S3	Direction
S4	Temperature
S5	Pyranometer
S6	Pressure
S7	Humidity

3. METHODS

3.1. Substitution Methods

Substitution methods assign mean and median values for each variable instead of the missing value. In order, for the mean values, to be used and substituted to missing values, these missing values should be within the context of MCAR [1]. This method is used for the variables which show a normal distribution. The most important aspect of this method is its simplicity and applicability [4].

3.2. Hot Deck Imputation

Hot Deck imputation is an important method which allows the missing value to be obtained from the data set without any further mathematical and statistical information [6]. The nearest k-neighbors Hot Deck method is one of the most preferred methods for recruiting the missing value. The distance between the lines of missing value and the complete lines are measured in this method; in other words, in the nearest neighbor Hot Deck imputation, k-nearest neighbors algorithm is used for obtaining the most proper line [4].

1) Data set is divided as incomplete data set and complete data set.

2) Let X_i be the data matrix specifying the complete set of data, and x_{ij} be i^{th} observation pertaining to j^{th} variable. And let Y_i be a data matrix specifying the incomplete data set and y_{ij} be i^{th} observation pertaining to j^{th} variable.

3) Euclides distances are computed for each line containing incomplete data set.

$$\text{Euclid(d)} = \sqrt{\sum_{j=1}^{n} (x_{ij} - y_{kj})^2} \quad (1)$$

After distances are determined based on the numbers of the nearest k number, an appropriate complete line is found and the missing value can be obtained for each incomplete data set[5].

3.3. Random Hot Deck Imputation

The procedure is composed of two stages. First of all, the data set is divided into two sets with some missing values and with some data complete. Then a line is randomly chosen from the complete lines in the data set in order to estimate any line having missing value [6].

4. THE COMPARISON OF PERFORMANCES

4.1. Standard Error

Standard error is used to describe the difference between observations estimated and observations in reality [12]. The real observation values (X_i) and the estimated ones ($\hat{X}_i$) in terms of Standard error are given as the following,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \hat{X}_i)^2}{n}} \quad (2)$$

4.2. Correlation Coefficient

The Correlation coefficients is between real values (X_i) and the estimated observation values ($\hat{X}_i$), and is given as follows,

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(\hat{X}_i - \overline{\hat{X}})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(\hat{X}_i - \overline{\hat{X}})^2}} \quad (3)$$

5. RESULTS

In this study, a properly developed C# software program was used. The missing values were randomly constituted in different proportions from the real data set. Each of the 7 variables in the data set together with the number of missing value and proportions (each variable has 4464 observations) were given in Table 2.

Table 2. Number of missing values for each variable and their percentages.

Variable	S1	S2	S3	S4	S5	S6	S7
Numbers of the missing value	218	340	423	499	614	377	457
Proportion (%)	4.8	7.6	9.47	11.1	13.7	8.4	10.2

Whereas the missing values in Table 1 were less obtained for S1 variable, they were the most for S5 variable. Correlation coefficients were used in order to determine to what extent the missing value were close to real data. The missing values obtained by using three methods Hot Deck, random Hot Deck, and substitution methods and real values based on the correlation coefficient were given in Table 3.

Table 3. Correlations between the real data set and those obtained by Hot Deck, random Hot Deck and substitution methods.

Variable	Hot Deck	Random Hot Deck	Substitution methods (mean)	Substitution methods (median)
S1	0.984	0.955	0.974	0.899
S2	0.983	0.926	0.966	0.869
S3	0.942	0.944	0.959	0.955
S4	0.963	0.939	0.940	0.940
S5	0.966	0.901	0.928	0.904
S6	0.969	0.940	0.956	0.954
S7	0.967	0.947	0.946	0.946

In Table 3, a strong relation was found between the values of the real data set and those obtained by Hot Deck, random Hot Deck and substitution methods. Hot Deck method proved to be the most exact one for obtaining the missing value closest to real ones.

Values of the Standard error and mean pertaining to the variables in missing values obtained by real data set and Hot Deck, random Hot Deck and substitution methods were given in Table 4.

Table 4. The Standard deviation and mean values pertaining to the data set obtained by Hot Deck, random Hot Deck and substitution methods through the real data set.

Variable	statistics	Hot Deck	Random Hot Deck	Substitution Methods (mean)	Substitution methods (median)	Real data set
S1	Mean	4.0799	3.9960	4.0787	4.2603	4.0809
	St. deviation	1.9495	1.9354	1.9005	2.0629	1.9522
S2	Mean	3.1584	3.0363	3.1603	3.3918	3.1564
	St. deviation	1.7470	1.7690	1.6911	1.8736	1.7500
S3	Mean	84.48	79.83	84.20	81.62	83.84
	St. deviation	78.12	75.56	74.10	74.52	77.30
S4	Mean	-2.4761	-2.3346	-2.4615	-2.4660	-2.4868
	St. deviation	4.5542	4.3136	4.2953	4.2953	4.5704
S5	Mean	106.57	95.34	107.14	92.41	107.45
	St. deviation	175.71	169.25	164.13	168.23	176.89
S6	Mean	101.76	101.76	101.76	101.77	101.76
	St. deviation	0.521	0.504	0.497	0.499	0.520
S7	Mean	72.402	72.477	72.407	72.345	72.359
	St. deviation	15.011	14.411	14.195	14.196	15.000

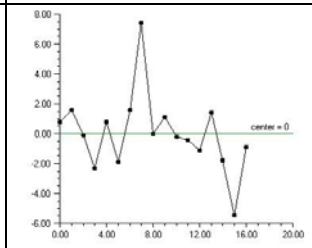
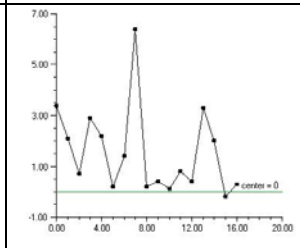
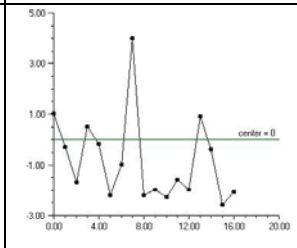
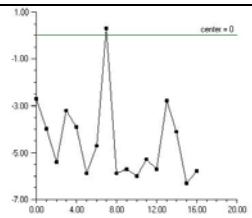
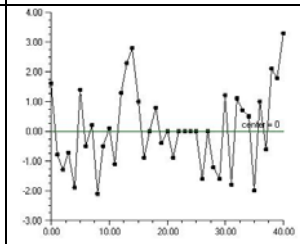
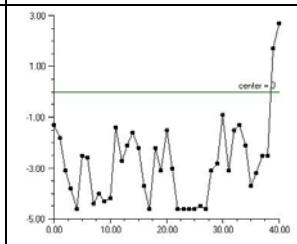
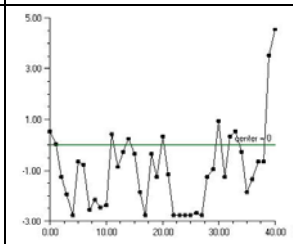
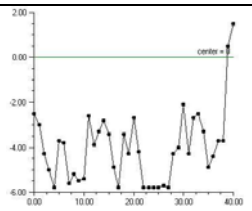
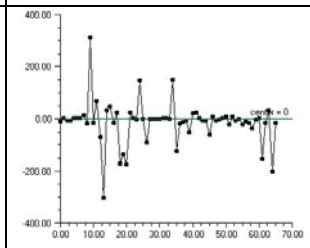
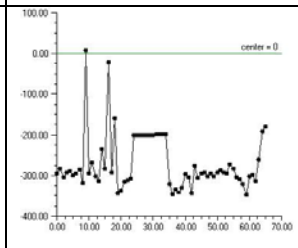
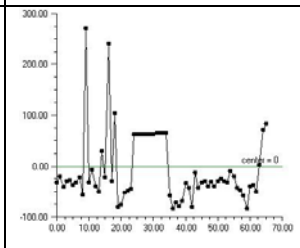
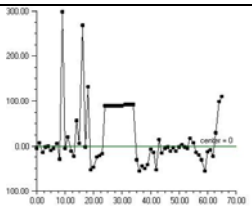
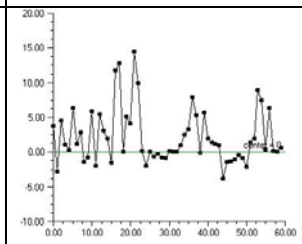
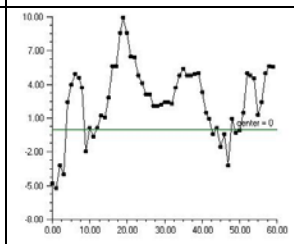
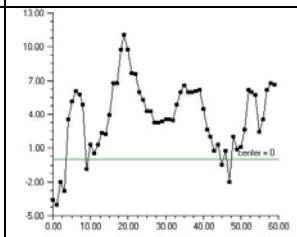
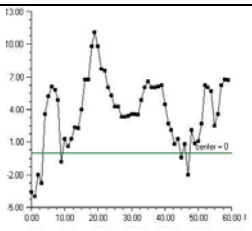
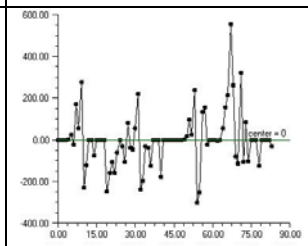
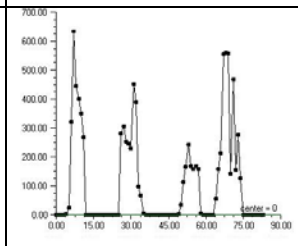
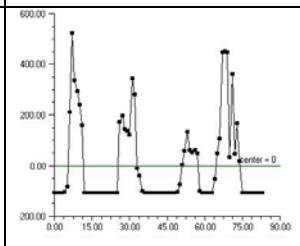
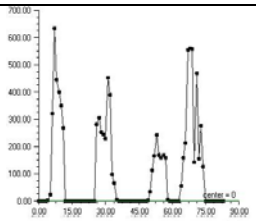
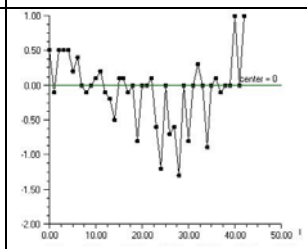
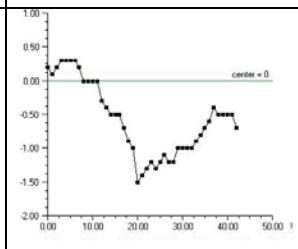
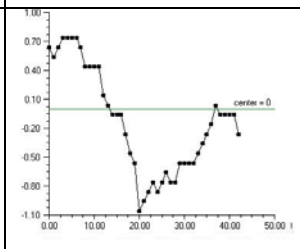
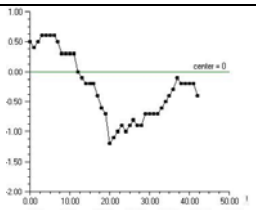
The mean and standard deviations pertaining to the variables of the data set obtained by Hot Deck method and those of the real data set were found to similar to each others. Those values obtained by random Hot Deck method were found to be smaller to those in the real set of data. The mean and standard deviations of those variables obtained by substitution methods were found to be greater than those of real data set. Therefore, Hot Deck method was found to be a much better tool in the estimation of the missing values. The Standard errors between the real data set and those obtained by four different methods were given in Table 5

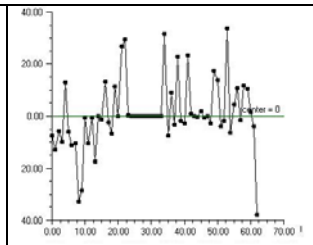
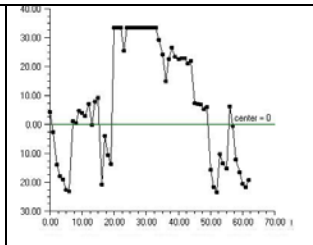
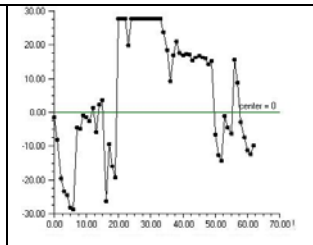
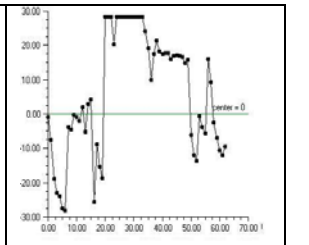
Table 5. The Standard errors between the real data set and those obtained by four different methods.

Model	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
Hot Deck	0.12	0.11	705.08	1.53	2126.04	0.02	14.98
Random Hot Deck	0.35	0.47	675.46	2.48	6135.16	0.03	23.63
Substitution Methods (mean)	0.2	0.2	485.11	2.44	5823.34	0.02	23.49
Substitution Methods (medain)	0.86	0.93	535.63	2.44	4562.44	0.02	23.33

When Table 5 is examined, the smallest standard deviation values for almost all of the variables are obtained through the Hot Deck method. Similarly, the graphics pertaining to the standard errors between the real observations and the missing ones by the four methods were given in Table 6.

Table 6. Graphics of the standard errors between the real data set and the missing values obtained by Hot Deck, random Hot Deck and substitution methods and those of the real data set.

	Hot Deck	Random Hot Deck	Mean Substitution	Median Substitution
S1				
S2				
S3				
S4				
S5				
S6				

S7				

Graphics were obtained using the standard error values between the real observations and the estimated missing ones. The closeness of the obtained standard errors to the zero point shows how the estimations of the missing values are near to those of the real ones. When graphics are examined, Hot Deck was found to be the most exact in terms of its closeness to the zero point.

6. CONCLUSION

The study software was developed using Microsoft C# programming language for the estimation of the missing value and the constitution of graphics. When the results are examined, Hot Deck method was found to yield the closest results to the real values compared to the other methods. Similarly, based on the estimations of the values obtained and those of the real values, the correlations, standard deviations, mean, median values and graphics compared between them, it would be much more tempting to say that Hot Deck method is found more effective and in concordance with the results of the previous literature [10,11].

REFERENCES

[1] Alan, O., Shaw, C., Lisa, H., "The comparative efficacy of imputation methods for missing data in structural equation modeling", ***European Journal of Operational Research,*** 151:53–79 (2003).

[2] Bal, C., Özdamar, K., "Solving The Missing Value Problem By Use Of Simulated Data Sets", ***Osmangazi Üniversitesi Tıp Fakültesi Dergisi***, 26(2):67-76 (2004).

[3] Jerez, J.M., Molina, I., Subirats, J.L., Franco, L., "Missing Data Imputation In Breast Cancer Prognosis", ***Processing of the 24th IASTED International Multi-Conference Boimedical Engineerring***, February 15-17, Innsbruck, Austria (2005)

[4] Joseph, L., Schafer, J., Graham, W., "Missing Data: Our View of the State of the Art", ***Psychological Methods***, 7: 147-177(2002).

[5] Juned, S., Thomas R.B., "Multiple imputation using an iterative hot-deck with distance-based donor selection", ***Statistics In Medicine,*** 27: 83-102 (2002).

[6] Kalton, G., Kish, L., "Some efficient random imputation methods", ***Commun. Statist.-Theor. Meth.,*** 13(16): 1919–1939 (1984).

[7] Mohamed, S., Marwala, T., "Neural Network Based Techniques for Estimating Missing Data in Databases", ***16th Annual Symposium of the Patten Recognition Association of South Africa, Langebaan,*** 27-32 (2005).

[8] Pelckmans, K., Brabanter, J.D., Suykens, J.A.K., Moor, B.D., "Handling missing values in support vector machine classifiers", ***Neural Networks,*** 18: 684-692 (2005).

[9] Stefano, M.I., Giuseppe, P., "Missing data imputation, matching and other applications of random recursive partitioning", ***Computational Statistics & Data Analysis***, 52: 773- 789 (2007).

[10] Wayne A.F., Jae, K.K., "Hot Deck Imputation for the response model", ***Proceedings of Statistics Canada Symposium*** (2001).

[11] Wayne, A.F., Jae, K.K., "Hot deck imputation for the response model", ***Statistics Canada***, 31(2):139–149 (2005).

[12] Yenduri, S., "Performance Evaluation of Imputation Methods for Incomplete Data Sets", ***International Journal of Software Engineering and Knowledge Engineering,*** 17:1-26 (2007).

APPENDIX

Developed software program using C# programming language.