ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

An Effective Method to Impute the Missing Values for Longitudinal Studies

P.Saravanan[1], S.Justin Samuel[2], V.Nirmalrani[3], S. Raaghavi[4], S.Vani[5]

Assistant Professor, Dept. of I.T., Sathyabama University, Chennai, Tamilnadu, India[1,3]

Professor, Dept. of I.T., Sathyabama University, Chennai, Tamilnadu, India[2]

Student, Dept. of I.T., Sathyabama University, Chennai, Tamilnadu, India[4,5]

ABSTRACT: Longitudinal studies are the concept in which the value for same variable changes periodically which when applied directly leads to misinterpretation of statistical data. Copymean is a new method proposed in order to impute the continuous missing values. It implements both cross sectional and longitudinal imputation methods on the data to obtain a rough imputed value which is efficient with assumption that the dataset is uniform. Here we propose a more efficient method firstly by applying a clustering technique on the dataset to minimize the data and then applying the copy mean algorithm to the sub-classified clusters iteratively. This alteration in the algorithm produces more accurate and appropriate imputed values which results in effective analysis of data.

KEYWORDS: Copymean, Imputation, longitudinal studies, clustering .

I. INTRODUCTION

A. Longitudinal data:

Longitudinal data are the variables whose values change periodically with time. These data keep changing from time to time. So when the value changes it has to get updated. Longitudinal reviews track similar individuals thus the differences seen in those individuals are more averse to be the result of social contrasts over time. Longitudinal studies subsequently roll out observing improvements more exact and are connected in different fields. Sometimes, information might go missing which has an impact on the database. So these missing values are to be considered and replaced properly. There have been numerous ways proposed to resolve this issue.

B. Imputation:

There are numerous methods for replacing the missing values like imputation, estimation, substitution, recovering actual values by contacting the person, and even deletion of the whole record. Imputation is the process of fetching the most closest value by various approaches like single imputation and various imputation and principled strategies. Under single imputation, there are various number of imputation strategies like hot deck, cold deck, mean substitution, relapse imputation. Multiple imputation incorporates three stages: imputation, analysis and combining the resultant outcomes. The list wise deletion removes the entire row that has missingness. This method is applicable in rare cases as every information is necessary. Other methods to recover data are pattern mixture and latent class models that check the similarity of the missing parameter with the other data for estimating the nearest value. In this paper, imputation process is applied to retrieve the missing value.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

II. PROBLEM STATEMENT

A. Occurrence of missing values:

Missing data can occur due to various reasons: It can occur

if there is no response from the users' side or can be caused by researcher or due dropouts. These missing data are classified into three types. They are

1) Missing Completely At Random (MCAR):

The missing data is neither related to the existing observed patters nor the missing patterns. This kind of missingness happens rarely and but if missingness is being created, then this type of missingness can be met.

2) Missing At Random (MAR):

The data whose patterns are related to observed data but not missing data. The values of the missing data can be predicted by analyzing the existing values of the other parameters. There should be a relationship or a pattern to know the similarity of the data among the parameters.

3) Missing Not At Random (MNAR):

The data whose arrangement relate to the missing patterns and depend on the missing data. This data occurs in rare cases and pattern mixture models is one of the solutions to find this type of data.

B. Quality of imputed data

The missing data can be resolved using the different approaches that are in existence. The crucial part of replacing this data is checking how close this value is to the true value. This can be known by finding the deviation constants.

The deviation can be defined as the difference between the true value and the imputed value. So if the deviation value is less, the imputed data is less deviated from the true value. There are various methods to find deviation like Root mean square deviation, Mean Absolute Deviation, Mean Square Error and Bias. These methods van be used to find the quality of imputation.

III. EXISTING SYSTEM

There are numerous methods that perform imputation by different algorithms. These methods are classified into cross sectional imputation and longitudinal imputation. The cross sectional imputation is the process of considering the data to be imputed for the subjects in a fixed time (for a year). the longitudinal imputation is imputing data for subject where the subject remains constant and is checked for a period of time (a series of years). the CopyMean method comes under both cross sectional and longitudinal methods.

The existing system makes use of the Copymean under cross-sectional and longitudinal imputations together. Initially, the longitudinal method LOCF (Last Occurrence Carried Forward) is used to obtain a value by considering the previous non missing value. Then, the mean value of the subjects is considered to refine the first approximated value so as to ensure that imputed value equals to mean value. The quality of imputation is calculated by measuring the deviation of imputed value from the true value. As a result, it is observed that the deviation value is least for Copymean algorithm when compared to the other conventional methods. CopyMean LOCF is efficient than other Copymean methods. The other methods, CopyMean global, local, and bisector were not efficient performance-wise because these methods depend highly on trajectory values hence making Copymean LOCF reliable.

Disadvantages of existing Copymean:

- Values are not accurately imputed..
- Method can be applied only on uniform dataset.
- This method is not suitable for static data.
- It has not been applied for real-time data.
- No clustering technique was employed which means it is not iterative.

IV. PROPOSED SYSTEM

The purpose of this paper is to focus on the Copymean imputation and the accuracy rates generated from the deviation resilts. The proposed method is efficient and has more accurate values when compared to the existing methods. Here, we

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

illustrate the influence of the missing data on the results of the longitudinal analysis. The missingness of data can be any of these missing types: MCAR, MAR or MNAR. We have proposed a more efficient method of imputation of missing value by initially clustering in order to filter the vast data. First, the data is clustered iteratively by using k means clustering technique. The iteration is done until the convergence is attained. Then the Copymean LOCF algorithm is applied to the sub-classified clusters that are obtained as a result of clustering. The CopyMean is already proved to be the efficient imputation algorithm. It was already proposed to combine the LOCF to the CopyMean. This alteration in the algorithm produces more accurate and appropriate imputed values which results in effective real-time data analysis. As a result of clustering, the data can be decreased and computation time is reduced. Now when the CopyMean algorithm is applied, it performs and imputes data with less divergence.

This method generates more accurate values as the data is minimized due to clustering. It can be applied on uniform datasets. It is performed on all types of missingness like MCAR,MAR and MNAR. This method proves to be efficient when compared to the other 12 conventional methods as the deviation constants are least when compared to the other results.

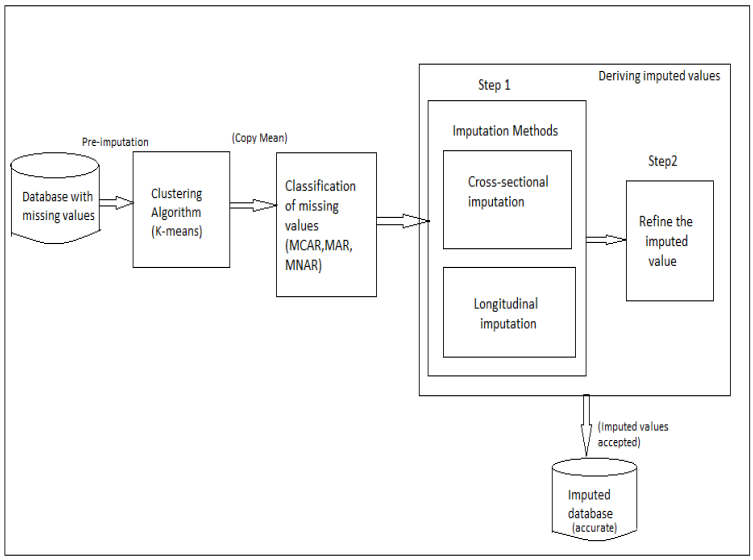

Figure 1

The above is an architectural representation of the whole process to be followed. The missingness of the data is detected and the tpre of missingness is known. Then Clustering is applied to minimize the data by reducing it to similar clusters. Then the existing Copymean algorithm is applied to this minimized data I order to attain a better accuracy.

V. WORKING PRINCIPLE

The Copymean method is followed by the clustering process. The whole method can be described as series of processes. The dataset is taken and it is a series real time dataset. The null values are created by considering some values as missing. This is done to check the quality of the imputed values. The clustering process is now applied which clusters and minimizes the data of similar groups together. This process is done so that the computational speed can be increased and the data retrieved can be more accurate. Then the Copymean algorithm is applied to the resultant clusters of data. This imputation is now compared to the true value by the deviation results and then the accuracy can be determined.

The above method is explained as a step by step process.

A. Dataset acquisition:

Any real time dataset can be acquired through various means like research conduct or through the datasets made available. In this paper the method has been applied on two datasets to check the performance and accuracy rates. The complete dataset with no missing values is taken and then missingness was created considering the parameters and the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

type of missingness.

B. Clustering the dataset:

The dataset is now clustered based on the similarity it possesses and is grouped based on similarly to get the minimized data in the form of sub clusters. The k means clustering technique has been employed to group the similar data. k-means clustering plans to segment n values into k groups where every observed value is grouped into a cluster where it is closest to the data center, filling in as a model of the cluster. This outcomes in a dividing of the data space. the data which is closest (euclidean distance) to the data center of the particular cluster belong to that respective cluster. So the data in a cluster are similar to each other and the data in the other clusters are dissimilar.

The squares of the observed value is calculated and the value whose mean is least is added into that particular cluster. So a data point is assigned into a cluster. Any data point can be a member of only a single cluster although it relates to more than one. The new means is calculated and this value is taken as the centroid for the particular cluster. The method generates k number of clusters with the k centers. The centers should be far from each other as the results depend upon the location. Now the data points are assigned into the clusters which are the nearest. After all the data points are assigned to the nearest cluster, the center is again assigned by calculating the mean of the clusters which leads to relocation of clusters. This process is repeated until there is no change in location of clusters. This clustering technique works fast and is understandable. The best results are attained when the data set used is diverse.

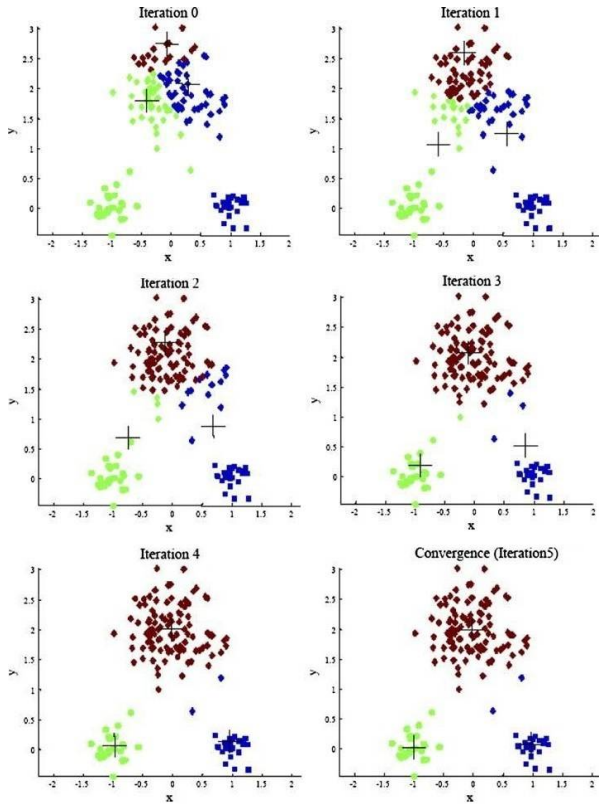

Figure 2

Figure 2 is a graphical representation of the iterations where the data center is relocated from time to time. When there is no change of location, convergence is attained.

C. Detection of missingness:

After the dataset with missing values has been created, the method has to detect the missingness in order to apply the Copymean algorithm. The null values are searched by the method by specifying the conditions. These missing values are now displayed from the whole dataset. To get the missing values, various technologies like ajax and jquery can be implemented. These perform the search for null values and group the retrieved missing values according to the users

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

request.

D. Copymean Application:

The Copymean LOCF [1] follows two steps: 1.It performs the LOCF method by taking the last found non missing value.2.Then the copymean method takes the previously found mean value of LOCF method and adds the variation. This resultant of addition is the new imputed value of the missing value. To compute this method, initially the LOCF method is applied to the whole trajectory and the last obtained value is considered as the resultant of LOCF method. Now, if more values are missing in the actual trajectory, then LOCF is applied again and the variation is calculated by subtracting the imputed trajectory from the average trajectory. Now this variation is added to the previously found LOCF value in order to get a similar mean value.

This method comes under both cross sectional and longitudinal implementations. The pregnant dataset is used on this method under MCAR category. This method is iterative and operates on all the values in the dataset. The accuracy level attained is more as the deviation is least when compared to all the methods. The Copymean method works successive to a method. It can be combined with various other methods to obtain better results.

$$y_{ij}^{CM} = y_{ij}^{LOCF} + AV_j. \quad (1)$$

In equation (1), Where y_{il}^{CM} is the value obtained by applying Copymean imputation method.

And y_{il}^{LOCF} is the result of last occurrence carried forward(LOCF) and AV_l is the average variation or the difference between $y_{.l}$ and $y_{.l}^{LOCF}$

$$AV_j = \overline{y_{.j}} - \overline{y_{.j}^{LOCF}} \quad (2)$$

.

In equation (2),

$y_{.l}$ is the value obtained by cross sectional imputation

And $y_{.l}^{LOCF}$ is the resultant of last occurrence carried forward.

The value $y_{.l}^{LOCF}$ can be calculated by considering the previous value before the missingness.

Hence the equations are applied to the existing data to obtain the outcome of Copymean

VI. DATASET AND IMPLEMENTATION

To implement the current method, two data sets have been acquired and imported. The data sets considered are complete. The missingness was created so that the comparison can be done between the true value and the imputed value.

A. Alcohol:

This dataset is a collection of wide range of attributes that contain details of alcohol consumption of teenage students and their parents' details and the effect of alcohol on their habitual life and their grades.

It contains many attributes like parents' employment and educational background as these factors indirectly influence their children in exposure to alcohol.

The dataset also includes other attributes like attendance records, participation in school activities, and their activities during free time.

The can be missing values generated as null and this missingness can be caused for any of the attributes randomly

The missing values were generated randomly and then algorithm was applied. This dataset contains data which is similar and relevant to each other i.e, one student's data can be similar to others'. Hence the algorithm works for the dataset by generating values for the missing attributes.

B. Air Quality:

The air quality data set contains the values of composition of various mixtures of gases in the air. It comprises the data with the records of date and time of the composition of these mixture of gases.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

This data set gives the information about the amount of gases and the increase in the pollution level. It describes about the variation of the gases on hourly basis for a particular number of days. It shows the estimates of the presence of various gases and their increase or decrease in the next hour.

The values of this dataset are closer to the previous and the next values as the variation between hour to hour is less. Hence the algorithm works for this dataset.

The missingness to this data can be created randomly as the values do not affect the change greatly. The Copymean algorithm is applied to the missing values and the resultant estimate is drawn.

VII. ACCURACY AND PERFORMANCE

The Copymean makes use of two datasets: alcohol and air quality datasets. Each dataset considers specific parameters by which the missing values are analyzed. In these datasets the missingness of data was created into numerous incomplete datasets and classified them into MCAR, MAR and MNAR. The datasets were implemented for both existing Copymean and the proposed new copymean. The accuracy of these methods can be calculated by deviation methods. Deviation is defined as the difference between the true value and the imputed value. The deviation and accuracy are inversely proportional i.e., if the deviation value is less, then the similarity between the true value and the imputed value is high. Hence the accuracy is high. There are various deviation methods : MAD, RMSD, MSE and Bias.

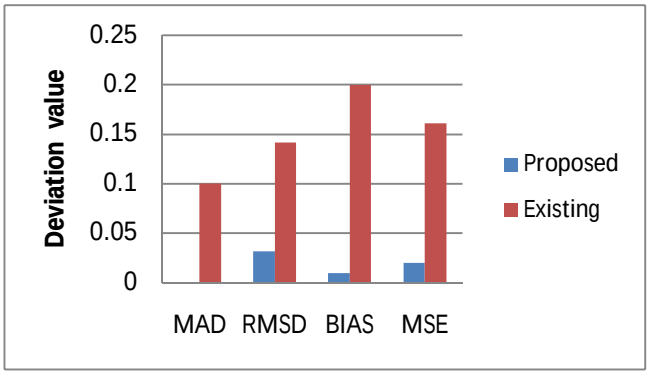

Figure 3

The figure 3 describes about the deviation results observed for the datasets. The deviation values are least for the air quality dataset. The deviation found is minute and negligible. All the deviation methods have been applied and the results to the data is represented as graph.

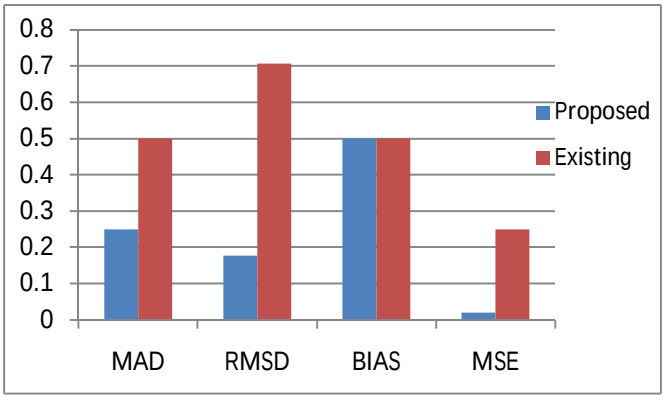

Figure 4

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

The figure 4 describes about the deviation results observed for the datasets. The deviation values are least for the alcohol dataset.

When the values are applied to the deviation formulas to the clustered data, CopyMean method is observed to have the least deviation and best results. Although the existing Copymean and linear interpolation method showed less deviation, it also showed less performance rate on some datasets and were not applicable on all types of missingness. Hence Copymean on the clustered data showed higher precision on all the datasets.

VIII. RESULTS AND CONCLUSION

From the above results, it is shown that imputational methods are better way to deal with the missing data than the principled methods as they cannot be applied for all types of missingness. Out of the imputational methods, the CopyMean is shown to have the better performance when compared to other imputational methods. This was proved by the least deviation values obtained for the CopyMean algorithm.

But it was also found more efficient when clustering technique was applied followed by Copymean method.

From the deviation results, we can state that the missing data imputation can be done efficiently through CopyMean technique when compared with the current methods. This is concluded because the data clustering limits the data to sub groups. Presently when the CopyMean calculation is applied to this minimized data, the values are imputed.

As the deviation of the imputed values is slightest in CopyMean calculation when compared with other methods, we can say that this strategy is more reliable and efficient in find missing values for the longitudinal data. In spite of the fact that there are different imputation methods proposed in the above literature survey, CopymeanLOCF was found to have improved performance considering the deviation results. Likewise the CopyMean works for both longitudinal and monotonous data on any types of missing values.

Unlike the principled techniques and other approaches to retrieve missing data, CopyMean works for MCAR, MAR and MNAR. Thus the conclusion can be drawn that of the present techniques for impuataion, CopyMean is reliable. It can prompt to better assumptions if more methods are combined to this algorithm to improve the precision.

The advantage of proposing the Copymean with clustering is that the data is minimized and hence reducing the values to be computed. By clustering the whole data can be reduced into parts and this categorization helps to perform imputation with ease.

IX. CONCLUSION AND FUTUREWORK

Copymean method is stochastic i.e., it is always followed by an initial method. In the existing system, CopymeanLOCF has been proposed where LOCF method is performed first and then copymean technique is applied. Likewise various other strategies and imputational methods can be included before copymean to attain more accurate values.

Clustering techniques can be applied to limit the datasets. The implementation can be done on this minimized data. This reduces the processing time and increases the computational speed.

Similarly many approaches like pattern mixture models and latent class models can be applied prior to this method to compare better results.

REFERENCES

[1] CopyMean: A new method to predict monotone missing values in longitudinal studies Christophe Genolini a,b,*, Amandine Lacombe a, René Écochard c,d, Fabien Subtil c,d
[2] J.M. Engels, P. Diehr, Imputation of missing longitudinal data: a comparison of methods, J. Clin. Epidemiol. 56 (10) (2003) 968–976.
[3] E. Dantan, C. Proust-Lima, L. Letenneur, H. Jacqmin-Gadda, Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts, Int. J. Biostat. 4 (1) (2008) 1–26.
[4] M.K. Olsen, K.M. Stechuchak, J.D. Edinger, C.S. Ulmer, R.F. Woolson, Move over LOCF: principled methods for handling missing data in sleep disorder trials, Sleep Med. 13 (2) (2012) 123–132.
[5] Y. Dong, C.Y. Joanne Peng, Principled missing data methods for researchers, Springerplus 2 (1) (2013) 1–17.
[6] C. Genolini, R. Écochard, H. Jacqmin-Gadda, Copy mean: a new method to impute intermittent missing values in longitudinal studies, Open J. Stat. 3 (2013) 26

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

[7] C. Genolini, B. Falissard, Kml: k-means for longitudinal data, Comput. Stat. 25 (2) (2010) 317–328.
[8] Missing data imputation in longitudinal cohort studies application of PLANN-ARD in breast cancer survival .
[9] Ana S. Fernandes2, Ian H. Jarman1, Terence A. Etchells1
[10] J. Twisk,W. de Vente, Attrition in longitudinal studies: how
[11] to deal with missing data, J. Clin. Epidemiol. 55 (4) (2002)329–337.
[12] W.J. Shih, H. Quan, Testing for treatment differences with dropouts present in clinical trials – a composite approach, Stat. Med. 16 (11) (1997) 1225–1239.
[13] M.S. Gold, P.M. Bentler, Treatments of missing data: a Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization, Struct. Equ.
[14] M.S. Gold, P.M. Bentler, Treatments of missing data: a Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization, Struct. Equ. Modeling 7 (3) (2000) 319–355
[15] N.M. Laird, Missing data in longitudinal studies, Stat. Med. 7 (1–2) (1988) 305–315.

[16] R. Ecochard, H. Boehringer, M. Rabilloud, H. Marret, Chronological aspects of ultrasonic, hormonal, and other indirect indices of ovulation, BJOG 108 (8) (2001) 822–829
[17] C. Genolini. LongitudinalData, R package version 2.3, 2014.
[18] C. Genolini, X. Alacoque, M. Sentenac, C. Arnaud, kml and kml3d: R packages to cluster longitudinal data, Journal of Statistical Software 65 (4) (2015) 1–34.
[19] C. Genolini, B. Falissard, Kml: a package to cluster longitudinal data, Comput. Methods Programs Biomed. 104 (3) (2011) e112–e121.