iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4111

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Load Balancing Based SRMI in Multicore Environment

K. Neeraja Raghava Lakshmi[1],G. Joji Varma[2] ,D. Tarun Surya Sai[3] ,K. Prem Prasad[4]
[1, 2, 3, 4,] *Epartment of IT, LBRCE*

Abstract: ***Sequential Regression Multiple Imputation (SRMI), by regression models missing variables are produced for each imputation. SRMI deals with the small samples of data. In dealing with the large data, the numbers of imputations are increased. The alternate method to deal with large number of imputations is PSRMI (Parallel sequential Multiple Imputation). In PSRMI the imputations are distributed among the cores based on the processor of the system. The imputations are distributed in two ways: Load Balancing and Non-Load balancing. In this paper we are using load balancing technique. Load Balancer is a virtual hardware that distributes the work equally among the cores. By using this technique, we are dealing with large data. By handling the number of imputations, it reduces the mean square error that results in reaching the original value.***
Keywords: SRMI ,PSRM ,Load Balancer, Imputation

I. INTRODUCTION

Now-a-days, many organizations are using large datasets like medical data, financial, census products, banking sector. There is a chance of missingness occurred in these datasets. PSRMI is the technique to overcome the missingness. In this technique the imputations are distributed equally among the cores based on the processor of the system. By using these technique numbers of imputations are increased, it reduces the mean square error that results in reaching the original value. In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Missing data can occur because of nonresponsive: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponsive than others: for example, items about private subjects such as income. Missingness occurs when participants drop out before the test ends and one or more measurements are missing. Data often are missing in research in economics, sociology, and political science because governments choose not to, or fail to, report critical statistics. Sometimes missing values are caused by the researcher-for example, when data collection is done improperly or mistakes are made in data entry. There are different mechanisms for the missingness, they are:Missing At Random(MAR),Missing Completely At Random(MCAR), Missing Not At Random(MNAR). Missing completely at random (MCAR) means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. Missing at Random, MAR, mean there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables. Missing Not at Random, MNAR, means there is a relationship between the propensity of a value to be missing and its values. If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR). Imputation is the process of replacing missing data with substituted values. In a multiple imputation, instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values. Application of the Multiple imputation requires three steps: imputation, analysis and pooling.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

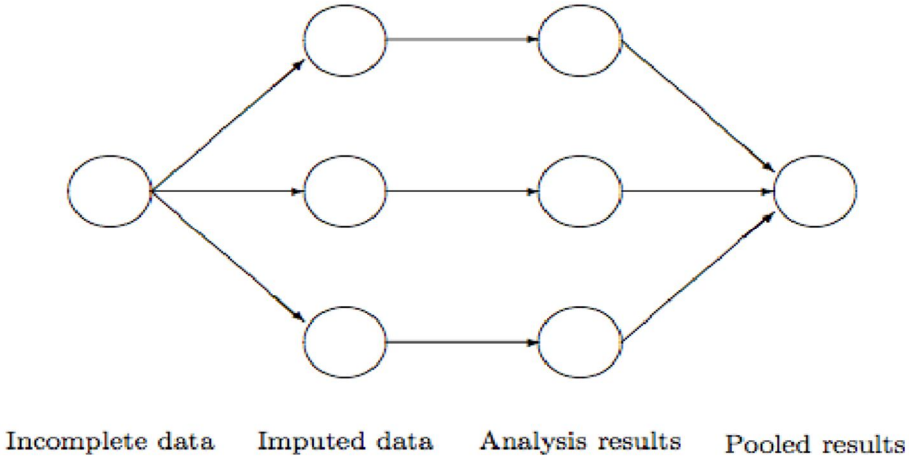


Fig 1.1 Represents Overview of Multiple Imputation

While measuring the performance we are considering the two parameters: Relative Bias, Relative Mean Square Error

A. System Architecture

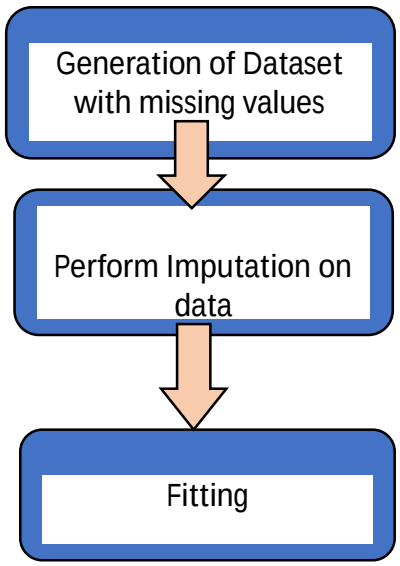


Fig 1.2 Represents Steps involved in PSRMI

A dataset with continuous data is invoked using functions which are available in R. Continuous Data is a data which can take any value (within a range). Then missing values are posed into the dataset based on the nature of missingness (MCAR, MAR, MNAR). m number of datasets are produced. Fitting distributions consists in finding a mathematical function which represents in a good way a statistical variable.

II. RELATED WORK

In this paper [1] multiple imputation method is used for handling the missing data. In this paper [2] they proposed two methodologies for handling missing data: ratio imputation and sequential regression multivariate imputation. They consider the NAPCS (North American Product Classification System) for product tabulations. In this paper [3] simple correction factor is applied to standard MI variance estimate, so that it can reduce the bias. In this Paper [4] different types of data like categorical, discrete, continuous are handled by trees, they can capture the important dependencies and interactions. In this paper [5]t-distribution is used as a default model with in SRMI for continuous data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

III. EXISTING SYSTEM

In previous systems they are used SRMI technique which is used for the small sample of data. Non-load balancing is used for handling the data, which takes more time to execute. In single imputation replaces the missing variable with a single value. By using single imputation, only single value is imputed which may not give the accurate values that is the relative mean square error is high.

IV. PROPOSED WORK

Missing observations are pervasive throughout observational research, especially in the social sciences. Despite multiple approaches to dealing adequately with missing data, many scholars still rely on list-wise deletion. The parallel sequential regression multivariate imputation (PSRMI, also known as chained equations or fully conditional specifications) is a popular approach for handling missing values in highly complex data structures with many types of variables, structural dependencies among the variables and bounds on possible imputation values. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes. A load balancer is a piece of hardware (or virtual hardware) that acts like a reverse proxy to distribute network and/or application traffic across different cores. A load balancer is used to improve the concurrent user capacity and overall reliability of applications. A load balancer helps to improve these by distributing the workload across multiple cores, decreasing the overall burden placed on each core.

V. IMPLEMENTATION

This system consists of different types of modules.

A. Data Preparation

A dataset with continuous data is invoked using functions which are available in R. Continuous Data is a data which can take any value (within a range). Then missing values are posed into the dataset based on the nature of missingness (MCAR, MAR, MNAR).

	lat	long	depth	mag	stations
1	-20.42	181.62	562	4.8	41
2	-20.62	181.03	650	4.2	15
3	-26.00	184.10	42	5.4	43
4	-17.97	181.66	626	4.1	19
5	-20.42	181.96	649	4.0	11
6	-19.68	184.31	195	4.0	12
7	-11.70	166.10	82	4.8	43
8	-28.11	181.93	194	4.4	15
9	-28.74	181.74	211	4.7	35
10	-17.47	179.59	622	4.3	19
11	-21.44	180.69	583	4.4	13
12	-12.26	167.00	249	4.6	16
13	-18.54	182.11	554	4.4	19
14	-21.00	181.66	600	4.4	10
15	-20.70	169.92	139	6.1	94
16	-15.94	184.95	306	4.3	11
17	-13.64	165.96	50	6.0	83
18	-17.83	181.50	590	4.5	21
19	-23.50	179.78	570	4.4	13
20	-22.63	180.31	598	4.4	18
21	-20.84	181.16	576	4.5	17
22	-10.98	166.32	211	4.2	12

owing 1 to 23 of 1,000 entries

	lat	long	depth	mag	stations
1	-20.42	181.62	562	4.8	NA
2	-20.62	181.03	NA	4.2	15
3	-26.00	184.10	42	5.4	43
4	-17.97	181.66	626	4.1	19
5	-20.42	NA	649	4.0	11
6	-19.68	184.31	195	4.0	NA
7	-11.70	166.10	82	4.8	43
8	-28.11	181.93	194	4.4	15
9	-28.74	181.74	211	4.7	35
10	-17.47	179.59	622	NA	19
11	-21.44	180.69	583	4.4	NA
12	-12.26	NA	249	4.6	16
13	-18.54	182.11	554	4.4	NA
14	-21.00	181.66	600	4.4	10
15	-20.70	169.92	NA	6.1	94
16	-15.94	184.95	306	4.3	11
17	-13.64	165.96	50	NA	83
18	-17.83	181.50	590	4.5	21
19	NA	179.78	570	4.4	13
20	-22.63	180.31	598	4.4	18
21	-20.84	181.16	NA	4.5	17
22	-10.98	166.32	211	4.2	12

ving 1 to 23 of 1,000 entries

B. Performing Imputation

m number of datasets are produced. Multiple imputations require three steps: imputation, analysis and pooling.

C. Fitting

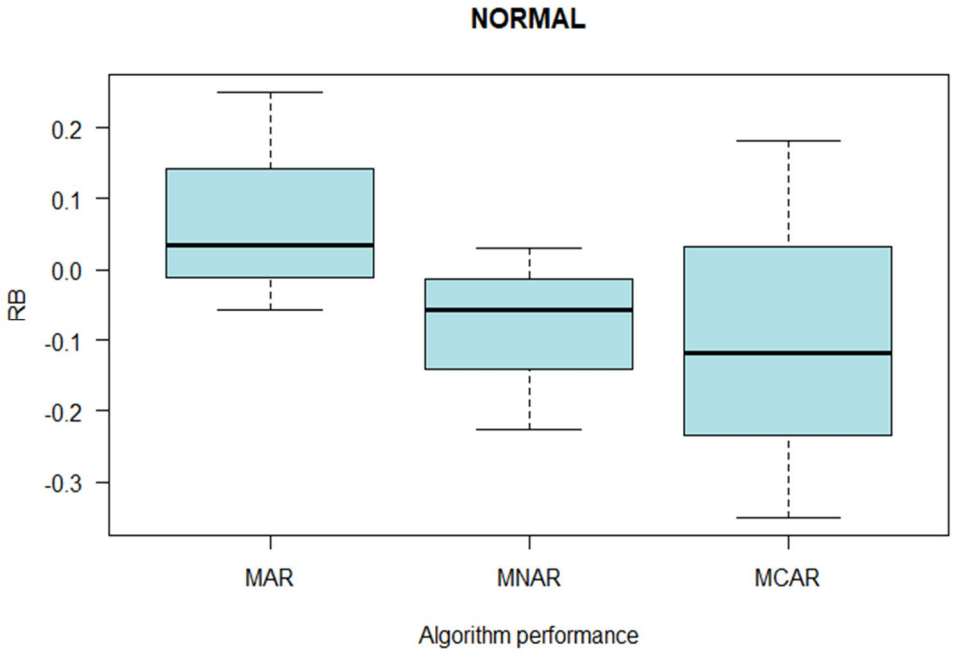
Fitting distributions consists in finding a mathematical function which represents in a good way a statistical variable.

D. Normalized Tables

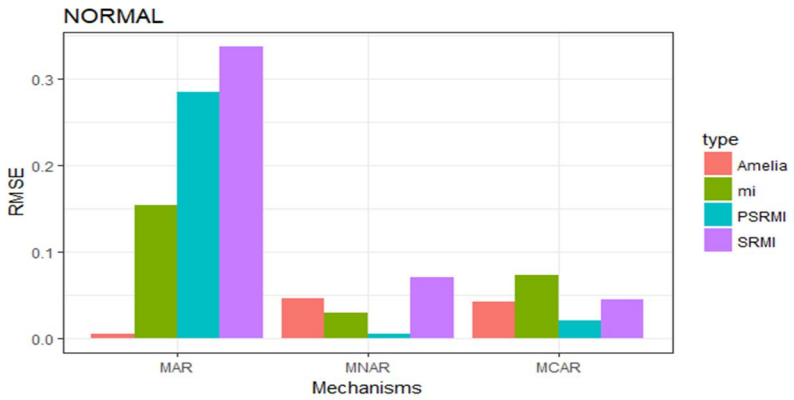
Imputation Method	MAR		MNAR		MCAR	
	RB	RMSE	RB	RMSE	RB	RMSE
M1	0.2501	0.15416333	0.0298	0.030166	0.18133	0.0735333
M2	-0.05726	0.00573033	-0.05683	0.045973	-0.11806	0.041902
M3	0.0345	0.33699	-0.2264	0.070306	-0.35006	0.0451784
M4	0.2533	0.28366667	-0.01006	0.005094	0.06733	0.0201

In MI[M1] method, produces high bias in MCAR and MAR and produces less bias in MNAR. The RMSE measure for these three mechanisms using MI provides efficient result in MNAR and MCAR. Amelia[M2] produces relatively less bias when compared with the above method. Amelia produces less bias in MCAR and high bias in MAR. SRMI [M3] produces the less bias in MCAR. By using SRMI[M3] the relative error is lower for MCAR and MNAR. In PSRMI the RB value is less in MCAR and MNAR. The relative error is less in case of MNAR. The bias in MAR is greater than MNAR by 80.2% and MAR greater than MCAR by 89.1%. PSRMI provides efficient result in MNAR and MCAR. PSRMI method gives relatively less bias and less error compared with other methods. In that MCAR and MNAR gives efficient results when compared with MAR mechanism in Normal Copula.

E. Plotting



The figure represents the Relative Bias values for the four methods in three different mechanisms (MCAR, MAR, MNAR)in normal copula. In MCAR and MNAR mechanisms the PSRMI method has performed well. Each method gives relatively similar bias in MNAR and MCAR.In normal MCAR has performed well.



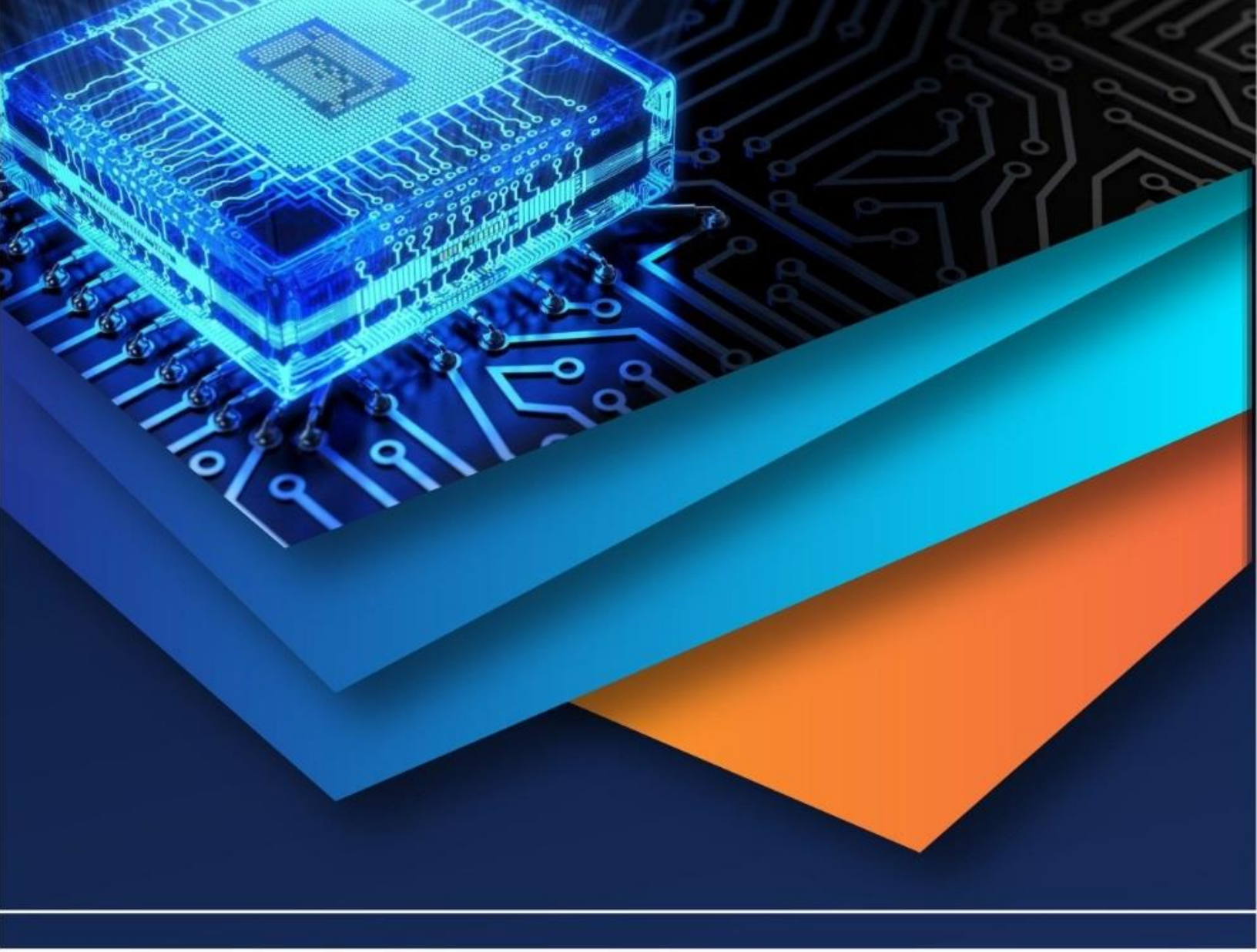
The figure represents the Relative Error values for the four methods explained above namely Amelia, Mice, SRMI and PSRMI in normal copula. In MAR mechanism the PSRMI method has performed well. MI[M1] method, produces high error in MCAR and MAR and produces less error in MNAR. The RMSE measure for these three mechanisms using MI provides efficient result in MNAR and MCAR. Amelia[M2] produces relatively less bias when compared with the above method. Amelia produces less error in MCAR and high error in MAR. SRMI [M3] produces the less error in MCAR. By using SRMI[M3] the relative error is lower for MCAR and MNAR. In PSRMI the Error value is less in MCAR and MNAR. The relative error is less in case of MNAR.

VI. CONCLUSION

PSRMI works effectively and give accurate values when compared to other techniques. PSRMI can handle with the large data like categorical, discrete, continuous. PSRMI reduces the Means square error which leads to original value. Load balancing will reduce the time of execution. Multiple Imputation can replace with m values.

REFERENCES

[1] Use of Multiple Imputationto Correct for BiasinLung Cancer Incidence Trends by Histologic Subtype Mandi Yu1, Eric J. Feuer1, Kathleen A. Cronin1, and Neil E. Caporaso2
[2] Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products1. Maria M Garcia, Darcy Steeg Morris, and L. Kaili Diamond US Census Bureau, 4600 Silver Hill Rd., Washington D.C., 20233
[3] A Comparison of Approximate Bayesian Bootstrap and Weighted Sequential Hot Deck for Multiple Imputation
[4] Tree-based prediction on incomplete data using imputation or surrogate decision Holger Cevallos Valdiviezo1 and Stefan Van Aelst2,11 Ghent University, Department of Applied Mathematics, Computer Science and Statistics,Krijgslaan 281 S9, Gent, Belgium2 KU Leuven, Department of Mathematics, Section of Statistics, Celestijnenlaan 200B B-3001, Leuven, Belgiu
[5] A Robust Model for Use in Sequential Regression Multiple Imputation M. J. von Maltitz, A. J. van der Merwe February 9, 2015
[6] Investigating the Performance of a Variation of Multiple Correspondence Analysis for Multiple Imputation in Categorical Data Sets Johané Nienkemper-Swanepoel Stellenbosch University, South Africa Michael J von Maltitz University of the Free State, South Africa



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY