A picture containing drawing

Description automatically generated

**DATA MINING**

**Project – C-20: Multivariate Imputation by Chained Equation (MICE)**

Submitted to:

Prof. Dr.Roozbeh Razavi Far

Submitted by:

Sindhu Leela Sishitha Pilli (110060987)

Ankur Patel (110060676)

**INDEX**

[1. Introduction 3](#_Toc99926224)

[2. Synopsis 3](#_Toc99926225)

2.1 Positive impact of using Multivariate Imputations

2.2 MICE Procedure

[3. Complexity analysis 5](#_Toc99926227)

[4. Implementation 5](#_Toc99926228)

[4.1 Libraries & its versions 5](#_Toc99926229)

[5. Experimental results 6](#_Toc99926230)

[5.1 Experimental Setting 6](#_Toc99926231)

[5.2 Results Analysis 6](#_Toc99926232)

[6. Conclusion 8](#_Toc99926233)

[7. References 9](#_Toc99926234)

# **Introduction**

Multiple Imputation with Chained Equations is a reliable and informative approach for coping with missing data in datasets. Through an iterative succession of prediction models, the process (imputes) missing data in a dataset. Each given variable in the dataset is imputed using the other variables in the dataset in each iteration. These iterations should be repeated until convergence appears to have been achieved. MICE is intended to work with MAR data. MICE aims to integrate the appealing characteristics of two schemes: regression-based imputation and multiple imputations [1].

**2.Synopsis**

Multivariate imputation by chained equations (MICE) is a Fully Conditional Specification-based imputation approach in which distinct models impute incomplete attributes. MICE has the unusual property of being able to handle diverse variable types (continuous, binary, unordered categorical, ordered categorical) by constructing alternative MICE algorithms [2]. As a result, each characteristic is modelled based on its distribution; for example, binary or categorical variables are modelled using logistic regression, while continuous variables are modelled using linear regression [2]. The modelled attribute represents the dependent variable in the regression models, whereas the remaining characteristics represent the independent variables. Because the MICE method assumes that missing values are MAR, its application in a dataset with missing values that are not MAR may result in biased imputations [3].

* 1. **Positive impact of using Multivariate Imputations**

Two methods are majorly employed in Multivariate Imputations. One, instances with complete data are utilised. The other, substituting missing values with mean, mode, median, or constant value. Typically, removing missing samples or restoring missing values with basic approaches produces bias in later dataset analyses. MICE's use of multivariate imputation decreases bias in the feature selection process [4]. MICE employ a logistic regression algorithm/log loss with a sigmoid function to impute null values in discrete data and linear regression to impute null values in continuous data in N iterations to get more accurate and consistent results by including the limiting error function (MSE).

The MAR assumption would imply that the likelihood of a certain variable being missing is purely influenced by the observed values, and that whether someone's income is missing is unrelated to their (unobserved) value.

* 1. Mice Procedure

The chained equation procedure is divided into four fundamental phases [4]

**Step 1**: A basic imputation, such as the mean, is used to impute each missing value. Mean imputations can be thought of as "placeholders" [4].

**Step 2**: Imputations for one variable (var) are reset to missing [4].

**Step 3**: In Step 2, observed values from variable "var" are regressed against other variables in the imputation model, which may or may not contain all variables in the dataset [4]. In other words, the dependent variable in a regression model is "var," whereas all other variables are independent variables. In general, these regression models are based on the same assumptions that would be used when executing a linear or logistic regression model in the real world.

**Step 4**: We then substitute missing values for "var" with predictions (imputations) based on our regression model 45], implying that when "var" is used as an independent variable, we will utilise both observed and imputed values in the regression models for other variables [4].

**Step 5**: After identifying each missing variable, repeat steps 2-4 [4]. One cycle is formed by iterating through each variable. After one cycle, all missing values are converted to predictions based on regressions that represent data connections [4].

**Step 6**: Repetition of stages 2–4, with each cycle updating the imputations [4].

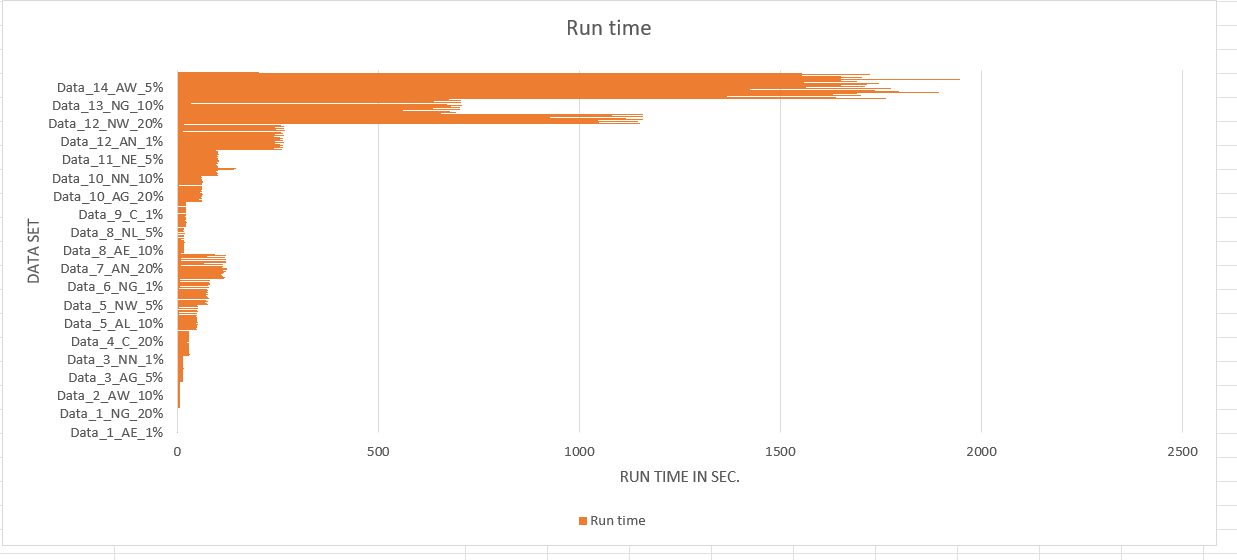
# **3.Complexity analysis**

Complexity analysis of any algorithm talks about how an algorithm scales with respect to the size of it input. For evaluating this, Big-o-notation is used. The syntax is represented as O(N). Consider N number of input variables available. Within the algorithm, there are nested loops. Our array has N items, our outer loop iterates N times, and our inner loop iterates N times for each outer loop iteration, giving us N2.

O (log N) is thought to be relatively efficient. The time required grows with the amount of the data collection, but not proportionally. This means that on smaller datasets, the process takes longer per item than on bigger ones.

For the current algorithm, our complexity analysis is O(N2\*logN) = O(N2logN)

The below graph shows a logarithmic inclusion of run time.



**4. Implementation**

The following instructions will help you run the code on Visual Studio(VS) Code: Simply copy and paste the entire code into VS, select the Excel files to target and it’s all set to go.

Install the specific versions of the following libraries depending on whether you are using Windows, Ubuntu, or Mac OS.

Run the file like any other Python program, e.g., python filename.py

## **4.1 Libraries & its versions**

et-xmlfile==1.1.0

joblib==1.1.0

numpy==1.22.3

openpyxl==3.0.9

pandas==1.4.2

python-dateutil==2.8.2

pytz==2022.1

scikit-learn==1.0.2

scipy==1.8.0

six==1.16.0

sklearn==0.0

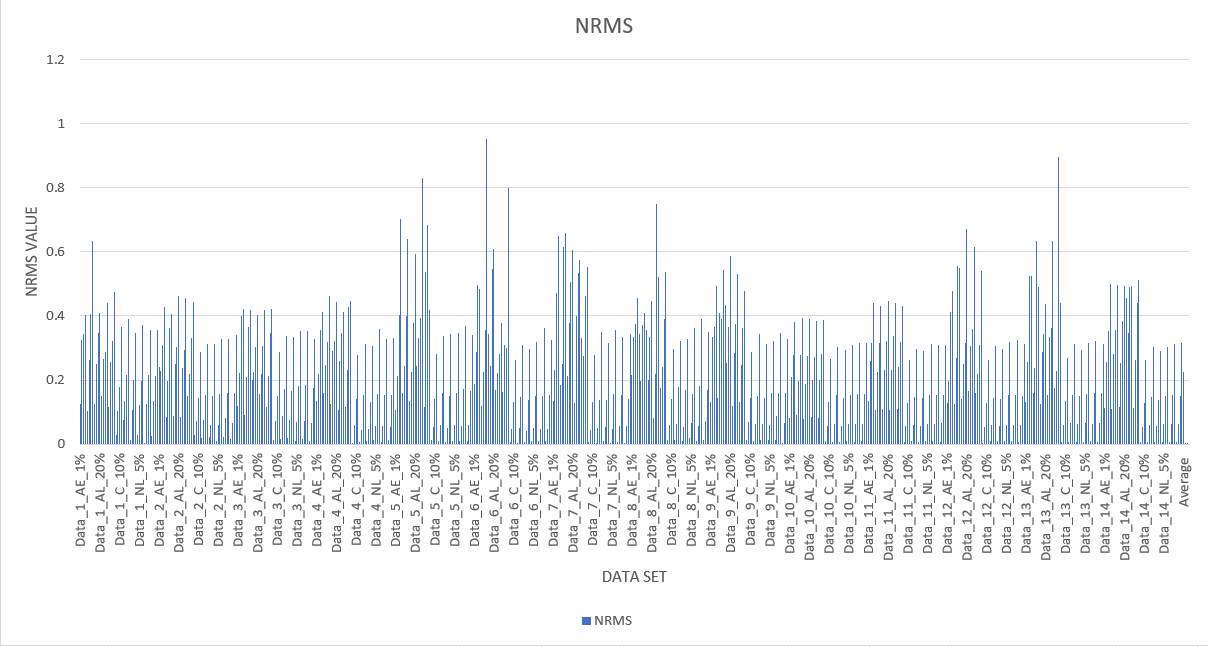
threadpoolctl==3.1.0

**5.Experimental results**

## **5.1 Experimental Setting**

## We experimented with many Datasets, each with a different number of iterations and tolerance rates. We discovered that there was no significant change in NRMSE and AE values when the number of iterations was kept at max and the maximum tolerance rate was lowered.

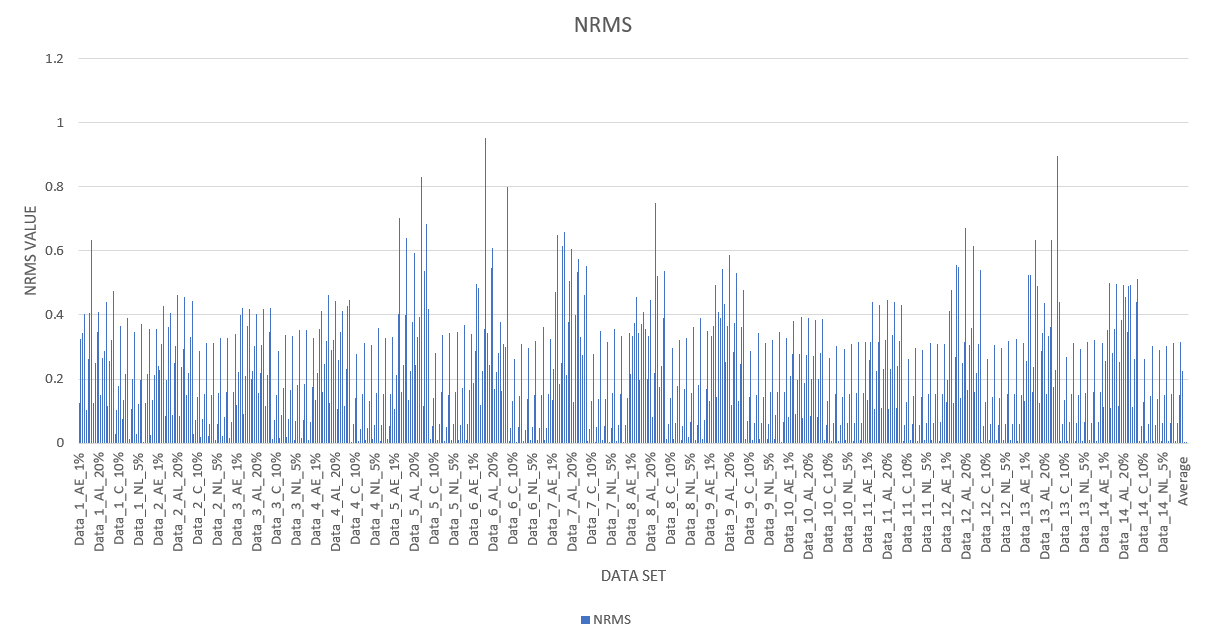
## As a result, limiting iteration and tolerance rates to a minimum saves execution time. For most datasets, we set the number of iterations to 100 and the tolerance rate to 0.01e-10.5.2 Results Analysis



## **5.2 Experimental analysis**

We assessed the impact of multivariate and multiple imputations on datasets containing categorical, numerical, and mixed (categorical and numerical) variables using the MICE method. To evaluate the technique, we examined datasets with varied frequencies of missing values.

The Histogram graph below indicates that a lower NRMSE value translates in better imputed data accuracy. Furthermore, when the missingness of the data grows, so does the NRMSE value. As a result, the lesser the missingness, the higher the accuracy of the imputed data.



Furthermore, the graph shows that, of all the incomplete numerical datasets, the dataset has the lowest NRMSE value (0.002725386), indicating that imputed data is more accurate than other datasets.

# **6. Conclusion**

The chained equation's multivariate imputation is a useful solution for the single imputation problem. Although the MICE technique is a sound strategy for dealing with missing data, several difficulties and limits must be acknowledged [4]. While MICE has several advantages over other missing data strategies in terms of flexibility, one major downside is that MICE lacks the theoretical foundation that other imputation systems have [4].

**7.References**

[1] D.Yaohui, and A.Ross, “A Comparison of Imputation Methods for Handling Missing Scores in Biometric Fusion.” Pattern Recognition, vol. 45, no. 3, Mar. 2012, pp. 919–33, doi:10.1016/j.patcog.2011.08.002, [Online].Available: <https://www.cse.msu.edu/~rossarun/pubs/DingRossMissingData_PR2012.pdf> [Accessed: Mar. 30, 2022].

[2] D.Bouhlila, and F.Sellaouti, “Multiple Imputation Using Chained Equations for Missing Data in TIMSS: A Case Study.” Large-Scale Assessments in Education, vol. 1, no. 1, Sept. 2013,p.4.BioMed Central, doi:10.1186/2196-0739-1-4, [Online]. Available: <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/2196-0739-1-4> [Accessed: Mar. 30, 2022].

[3] S.Buuren, K.Oudshoorn. “MICE: Multivariate Imputation by Chained,” *JSS J. Stat. Softw.,* 2011, [Online]. Available: <https://www.jstatsoft.org/article/view/v045i03> [Accessed: Mar. 30, 2022].

[4] M.Azur.,*et al*. “Multiple Imputation by Chained Equations: What Is It and How Does It Work?” *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, Feb. 2011, pp. 40–49. *PubMed Central*, doi:[10.1002/mpr.329](https://doi.org/10.1002/mpr.329), [Online]. Available: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/mpr.329>  [Accessed: Mar. 30, 2022]