Original Research Article

Close-Knit-Regression: An Efficient Technique In Estimating Missing Completely at Random Data

.

ABSTRACT

The study aimed at using the Close-Knit Regression (CKR) technique to approximate values absent because of the missing completely at random mechanism. Bivariate datasets were generated and simulated for MCAR mechanism at low (10%) and high (60%) rates. The CKR method was used and compared alongside other single imputation techniques like mean imputation, simple regression and K- Nearest Neighbors (K-NN). The difference between parameter estimates like mean, correlation coefficient (*r*), maximum, minimum and standard deviation which were gotten using predicted data and those using the original data as well as assessment of error rates like mean absolute error (MAE) and root mean square error (RMSE) were used as metrics in deciding the efficiency of the techniques. Results showed that the CKR technique was the best from those considered, with its estimated data having parameter estimates closer to that of the original whilst having the least error rates at 10% (MAE of 0.01 and RMSE of 0.047) and 60% (MAE of 0.021 and RMSE of 0.073) in comparison to other methods, CKR technique is a suitable single imputation technique which produces estimates close to the original data and parameters with low error rates when data are MCAR.

.

Keywords: Missing Completely at Random; Close Knit Regression; Mechanism; Parameter estimates; Mean Absolute Error; Root mean square error.

1. INTRODUCTION

The possession of high quality data is primarily important in research studies, a statistician, no matter his level of expertise can do from little to nothing without access to reliable information on the phenomena he wishes to asses.

It is in fact safe to say, that one can not depend on the results of any investigation if the data source is not verifiable. In the real world however, data collection is affected by so many factors, ranging from human error or apparatus failure to voluntary or involuntary non response or invalid answers by some participants and even loss of life[1].

While some of the aforementioned dynamics are mitigatable, most are not within the complete control of the researcher which makes avoiding them almost hopelessly inevitable [2],leading to unwanted errors, lack of consistency alongside redundancy and inadequacy in data sets. This in turn can heavily compromise the process and outcome of data analysis if not making it impossible to proceed in some cases.

When there are no values recorded in required information fields during research, it is referred to as *missing data*[3]. It is the lack of input, where input is needed. It can also be referred to as information that should have been present but isn't, for peculiar reasons [4]. According to McKnight et.al. [5] the causes of missing data can be usually traced to:
(a) The study participants, which entails errors on the part of subjects or their refusal to provide information for personal reasons (participant characteristics).
(b) The study design, having to do with the structure of the data collection methods and how its tedious and overbearing nature could discourage participants from providing complete data (design characteristics).
(c) The interaction of (a) and (b) above that has to do with the repercussions from the contact of study participants with design, an example of this is when some subjects in clinical trials are too sick to continue. There have also been cases of missing values due to the aforesaid reasons occurring in non-indigenous forms, they camouflage among genuine data making the task of spotting them a strenuous one.[6]

The course of action that led to missing values existing in a data set is referred to as the *mechanism of missing data* [7]. Little and Rubin [8] gave a deft classifying system of missing values basing mainly on their probabilities. When the probability of a variable being missing is independent of all other variables (observed and unobserved) in the data set, the mechanism in place is Missing Completely at Random (MCAR), a good example is skipping of certain items on a questionnaire by respondents due to oversight. Sometimes, the probability of a variable being missing is dependent on other observed variables in the data, this defines as Missing at Random (MAR), for example, women might exclude their age response in the demographic section of a questionnaire for sociological reasons. The last mechanism is Not Missing at Random (NMAR) and this happens when the probability of missing value occurrence is dependent on both observed and unobserved variables, take for example data on the IQ scores with data missing for subjects with low IQ values. The lack or presence of constancy in the way data are missing is referred to as its *pattern.* A univariate pattern happens when values are absent for only one variable. When missing values are dependent on each other it is termed to have occurred in a monotonic pattern, arbitrary patterns occur in random fashion[5].

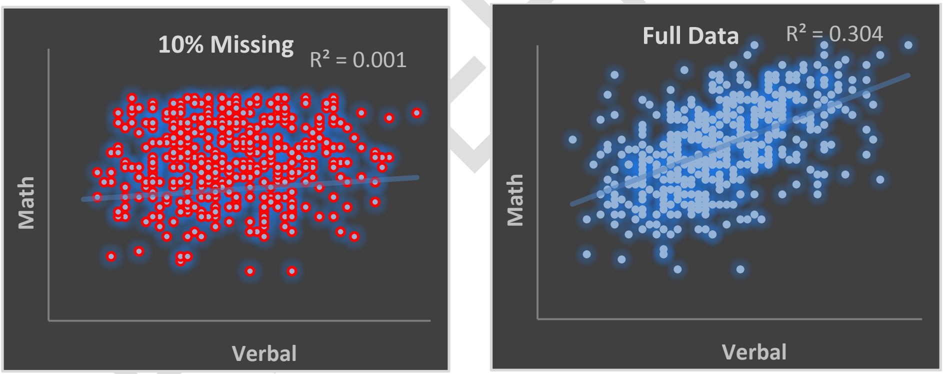

Fig 1. Showing the Effect of Missing Data on a Scatter Plot of Scores from a Math and Verbal IQ Test

Notwithstanding the advent of super computers with high end estimating powers in the 21st century, the problem of missing value estimation has continued to trouble researchers and scientists alike[9] .Its predominance in datasets if not addressed, being one of the many causes of bias when estimating parameters[10],hence weakening the statistical and empirical powers of estimators. There are a plethora of techniques for handling missing data ranging from complete/available case analysis to single imputation methods, likelihood based approaches and multiple imputation techniques[11]. Single imputation being one of the most flexible and general methods is easier and more direct than other techniques this in turn makes it more popular. Single imputation techniques however, tend to ignore uncertainty and almost always underestimates variance, like it was evident in the research of Paniagua et. al [12].

This thesis aimed to develop and apply the close-knit-regression (CKR) approach as a single imputation method, methods, investigate its advantages and disadvantages (if any) alongside three (3) other selected single imputation techniques in widespread use, which are mean imputation, simple linear regression and K-Nearest Neighbour (K-NN) . Which for a wider scope,it will make use of these methods tried on generated data which will be simulated for MCAR mechanism at low and high rates of 10% and 60% respectively under a univariate pattern.

Literatures surveyed on single imputation methods indicate that The CN2 and C4.5 algorithms are peharps the two most simplest of all imputation methods, they in general replace missing values with the mode from entries of the variable considered. A study by Grzymala-Busse and Hu [13] categorized them both as not very good estimators of missing values. These findings were also supported by [1] in later studies which compared the two aforesaid algorithms with more precise procedures like the K-NNI method. The mean imputation has been found to recurrently underestimate standard error of parameters [10], [14], [15].Simple regression and using conditional means were both deemed more effective method than mean imputation [16].

Given the number of repetitious cases of missing values post data collection, a good portion of statisticians have since proposed a variety of single imputation methods that handle such inconveniences The CKR was developed to make up for some of the shortcomings of other popular single imputation methods. The proposed CKR method is expected to not overly underestimate variance while providing more accurate estimates since the imputations are conditionally random, systematic and likely to be different for each missing point.

2. MATERIAL AND METHODS

Data simulations will be performed in R using the ampute function as proposed by Schouten et.al [17] which works with mice, vim and MASS packages. Continous defined datasets of one thousand observations (N=1000) will be generated**,** which will be composed of two fairly correlated variables (*V1, V2*) as most real world variables are, be aware that the covariance matrix should be semi definite. Summary of variables and conditions used in this study will be specified in Table 1 below.

Table 1. Summary of Variables and Conditions Used In This Study

Variables	**Correlation (*r*)**	**Missing mechanism**	**Missing Pattern**	**Distribution**	**Missing Rate**	**Techniques**
V1(Independent variable) and V2(Dependent Variable)	Fairly correlated (0.4)	MCAR	Univariate on dependent variable (V2)	Both are standard normal *V1, V2 ~ N* (0,1)	10% & 60%	Single Imputation : 1) Mean Imputation 2) K-NN 3)Regression Imputation 4)CKR *(Proposed Method*)

2.1 Data Set Simulation

After the data set generation is complete, ampute function has several other arguments which specify the nature of your missing data. First is the proportion, which in our study will vary from 10% to 60%. Next is the specification of missing mechanism which for our study will be of the univariate kind acting on the dependent variable.

Another important argument in the ampute function is the one that lets you select the frequency of missing-ness across the data, ampute divides original data into multiple subsets, where the number of subsets which has values in proportions that sum must equal one using a single value will suit the univariate pattern assigned earlier.

Specification of the mechanism to be MCAR is the next step after which assigning the weights which determines the relative missing-ness in the data set with respect to the variables, A weighted sum score uses a linear regression with coefficients assigned, it is of the form

$$wss_i = W_1 \cdot V1i + W_2 \cdot V2i \ldots (1)$$

where wssi is the weighted sum score of case i, V1i and V2i are the variable values of case i and W_1 and W_2 are the specified weights.

Keeping in mind that MCAR is completely random and the variables don’t influence its being missing, a zero weight is assigned to both variables. The last argument in ampute is not applicable to the MCAR mechanism.

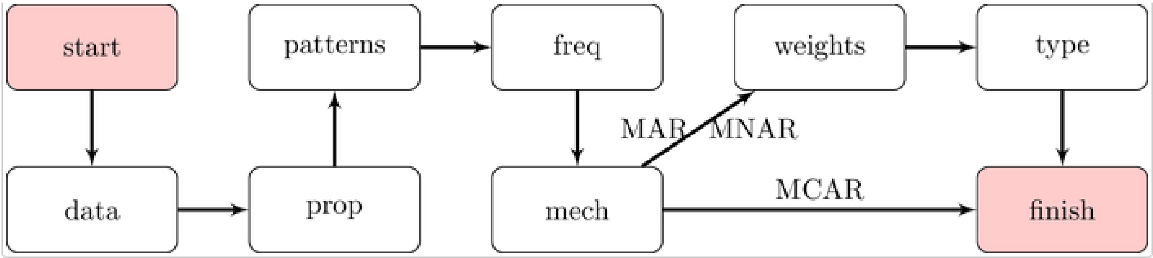

Fig 2. Flowchart showing steps in the 'Ampute' process(Adapted from Schouten et.al [17])

2.2 Techniques Considered In the Study

A total of four single imputation techniques were considered in this study, three (3) are already commonly used and the last is the method proposed, they are :

a)Mean Imputation
b)K-NN
c)Simple Regression
d) CKR (Proposed Method)

A brief description of these methods will be in focus.

2.2.1 Mean Imputation

The mean imputation is one of the most popularly known methods. It replace the missing values in a variable with the mean of all present values for continous data, while it replaces the missing values with the mode in discrete data. The disadvantages of the mean method is mainly on how it tends to underestimate variance by repeating values since the mean is a constant, correlation coefficient values are also stunted cause of the repititive nature of its outcome. Mathematically If x_{ij} of the k-th class Ck is missing, then it is replaced by

$$V2i = \sum_{i:\, v2i\epsilon C_k} \frac{v2_i}{n_k} \ldots(2)$$

2.2.2 K-Nearest Neighborhood (K-NN).

The K-NN method replaces the missing values by considering the given number of occurences that are most similar to the value of interest.It has numerous advantages, as it can be used for both qualitative and quantitative features in a data set, it doesn't make use of a predictive model too, the K-NN method also considers the correlation structure of the data. The first set back of this method is in the consideration of what distance function to use, it also requires a lot of time which is based on the choice of k. The procedure is as follows:
a) Given a data set V2, Divide V2 into two parts. Let $V2_{mis}$ be the set containing the instances in which at least one of the features is missing. The remaining instances with complete feature information form a set called $V2_{pres}$.
b) For each vector *V2* in $V2_{miss}$: Divide the instance vector into observed and missing parts as V2 = [$V2_{obs}$; $V2_{miss}$]. Calculate the distance between V2 and all the instance vectors from the set $V2_{pres}$. Use only those features in the instance vectors from the complete set $V2_{pres}$, which are observed in the vector V2.
c) Use the K closest instances vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the k-nearest neighborhood. The median could be used instead of the mean in cases of categorical data.

The K-NN takes into consideration the correlation structure of the data set and is so an improvement on using the mean.

2.2.3 Regression Method

This is usually used for univariate or monotone missing data pattern. The first step involves building a model from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data. The demerits of this method is usually the model estimated values are usually more artificial than the true values, also the technique could suffer from a lack of precision especially if there are no relationships among the values in the data set and the attribute with missing data, it is sometimes a tedious process too, since depending on the number of variables with missing data, so many models could be created .

Suppose that there are 2 variables *V1*, *V2* in a data set and missing data are uniformly or To impute the missing values for a variable, one first constructs a regression model using observed data on *V1* through *V2*.

$$V2 = \beta_o + \beta_1 V1 \ldots (3)$$

The regression model in above yields the estimated regression coefficients β_o, β_1 and the corresponding covariance matrix. Based on these results, one can impute one set of regression coefficient. from the sampling distributions of β. Next, the missing values in *V2* can be imputed by plugging β_o, β_1 into the above equation and adding a random error ε resulting in one complete data set.

2.2.4 Close-Knit Regression (CKR): Proposed Method

The proposed method combines certain aspects of the K-NN regression with simple linear regression, The close-knit-regression has two stages, first the close-knit sample-selection-stage where numerical values present in the incomplete data set that we think would give us the best estimate of missing data points are selected. Then the estimation stage where linear regression is applied to the selected sample and a model is built to use in interpolating (preferably) or extrapolating missing data points. It was built to handle univariate missing patterns.

Given two fairly correlated variables (V2,V1). Let V1 (v_{1i}'s) be the complete data set of the predictor variable, and V2(v_{2i}'s) the outcome variable with some missing values, for a univariate missing pattern in (V1,V2). To use the close-knit regression algorithm of V2 on V1 to estimate missing values in V2, we follow the steps below:

a) Sort the entire data set, by re-arranging the complete predictor variable V1 in ascending or descending order.

b) For a value say $V2_n$ missing in the outcome variable V2, compute all $|V1_n$-$V1_i|$'s, a set of absolute differences.

c) Say the smallest absolute difference is obtained at $V1_i$=$V1_a$

==> $|V1_n$-$V1_a|$ **<** all $|V1_n$-$V1_i|$'s for all values of *i* not equal to *a*.

And it is so that $V1_a$ has a corresponding non-missing value in V2 say $V2_{a.}$

form a set of closely knitted samples, C and add ($V1_a$,$V2_a$) as the first set of element, that is C= {($V1_a$,$V2_a$)},

d) i) If $V1_n$-$V1_a$>0 i.e $V1_n$>$V1_a$ then for the next entry $V1_b$ with a corresponding $V2_b$ value, search for values closest to $V1_n$ i.e the smallest $|V1_n$-$V1_b|$ where $V1_n$-$V1_b$ < 0 i.e $V1_n$ < $V1_b$.

ii) If on the other hand, $V1_n$-$V1_a$<0 i.e $V1_n$<$V1_a$ then for the next entry $V1_b$ with a corresponding $V1_b$ value, search for values closest to $V1_n$. i.e the smallest $|V1_n$-$V1_b|$ where $V1_n$-$V1_b$ > 0 i.e $V1_n$ >$V1_b$.

iii) If no such values exists as in i or ii, then for the next entry $V1_b$ with a corresponding $V2_b$ value, only search for values closest to $V1_n$ i.e the smallest $|V1_n$ -$V1_b|$.

e) In similar fashion, sets of bivariate entries (V1,V2) are added to the set C till a chosen number of elements which is the close knitted sample size (*n*) is reached.

n{C}=*n*

f) Simple-linear regression involving the elements of C is then performed to obtain coefficients, these are then used to estimate the missing data point $V2_{n.}$

g) The procedure is repeated till there are no missing data points in V2.

The logic behind this method is straightforward, once a missing data point is located in our outcome variable V2, find data points in the predictor variable V1 that are nearest to value that was supposed to have generated the missing point in V2. Then use a selected number of those points in V1 to build a model involving non missing points in V1 and V2 which will be used to give the best predictor equation of the missing point in V2. The method is expected to produce good parameter estimates while not inflating their standard errors.

2.3 Performance Measures

The indicators used to asses the precision of the missing data techniques relative to the complete data are the correlation coefficients (r), means, minimums, maximums, ranges, mean absolute errors and root mean square errors, they are described briefly below.

2.3.1 Comparison of Parameters

Firstly, the arithmetic mean of the complete data and imputed data will both be calculated and compared using the basic formula given by:

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{V}_i \ldots(4)$$

Where $\bar{Y}$ is the mean of the data in focus, n is the size, and $\hat{V}_i$ the data points. The mean will tell us about the comparative centrality of our datasets. Next, the correlation and standard deviation of the complete data and imputed data will also both be estimated and assessed comparatively using the Pearson correlation coefficient formula given by the two formulas respectively

$$r_{V2V1} = \frac{\sum_{i=1}^{n}(V1_i - \overline{V1})(V2_i - \overline{V2})}{ns_{v1}s_{v2}} \ldots(5) \qquad S_{v2} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{V2_i} - \overline{\mathrm{V2}})^2} \ldots(6)$$

Where for n data points, $V1_i$ *and* $V2_i$ are the values of both complete and estimated data points of V1 and V2, with means and standard deviations $\overline{V1}$/ $\overline{V2}$ and s_{v1}/s_{v2} respectively. The values of each of the correlation gotten from MDTs will be compared with that of the complete data. Contrasting the correlation coefficients and standard deviation will tell us about the spread and the strength as well as direction of the bivariate linear relationships and existing in the full and imputed data sets respectively.

The maximum and minimum values from the complete and estimated data points of the outcome variable Y in focus will gotten and there on, used to calculate the range to give us a quick sense of the spread.
Maximum value of $\hat{\mathrm{V}}_{\mathrm{i}}$ =max ($\hat{\mathrm{V}}_{\mathrm{i}}$) , Minimum value of $\hat{\mathrm{V}}_{\mathrm{i}}$ =min ($\hat{\mathrm{V}}_{\mathrm{i}}$) ,

Range $\hat{\mathrm{V}}_{\mathrm{i}}$ = max ($\hat{\mathrm{V}}_{\mathrm{i}}$) −min ($\hat{\mathrm{V}}_{\mathrm{i}}$).

2.3.2 Comparison of Errors

We here will be comparing the error arising from the differences in values between the complete simulated data and that estimated We will be using the Mean absolute error (MAE) and the Root mean square error (RMSE). To compute the MAE and RMSE, the difference between the estimated dataset points ($D_{est.}$) and complete data set points (D_{com}) will be used to get the MAE and RMSE which represents the sample standard deviation of the MAE [18].

$$MAE = \frac{\sum_{i=1}^{n}|D_{com} - D_{est}|}{n} \ldots(7) \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n}(D_{com} - D_{est})^2}{n}} \ldots(8)$$

3. RESULTS AND DISCUSSION

Results of data analysis after simulations are presented in this chapter, the techniques were applied to the datasets altered to suit the conditions given in Table 1.

3.1 Presentation of Results

Results are shown in terms of the proximity of the parameters estimated using techniques to that from the original dataset (Table 2 and Table 3) and then consideration was given to the error rates the parameters generated (Table 4 and Table 5).

3.1.1 Comparison of Parameter Estimates

The summary of statistics of the originally generated data before missing conditions were implemented showed that fair correlation between the variables V1 and V2 with, $r_{(V1,V2)}$ = 0.4, Our variable of concern was V2 where the minimum and maximum values were -0.98 and 8.67 resulting in a range of 9.65.. V2 also had a mean and standard deviation of -0.0001 and 1.001 respectively.

Results of Table 2, at 10% missing-ness for MCAR mechanism, CKR (Our proposed method) produced estimates with the best proximity to the full data with correlation coefficient of 0.39, mean of -0.003, minimum of -0.71, maximum of 8.3 and a range of 9.01. Results of simple linear regression were closely related to those of k-NN and Mean. The simple linear regression technique produced results with a correlation coefficient of 0.37, a mean of 0.001, minimum of -0.5, as well as a maximum and range of 8.2 and 8.7 respectively. The mean imputation technique produced data points with a correlation of 0.36, mean and range of 0.01 and 7.74 respectively while having a minimum of -0.44 and a maximum of 7.30. For K-NN imputation, the generated data points had a correlation coefficient of 0.36 with mean and a range of 0.03 and 8.03. The least value was -0.63 and the highest was 7.4.

Table 2. The Four Parameters Before and After Estimation from MCAR with Techniques at 10% rate

	MCAR @ 10%				
Parameter	Single Imputation Technique				
					Proposed Method
	Full Data	Simple Reg	Mean	K-NN	CKR
r	0.40	0.37	0.36	0.36	0.39
Mean	-0.0001	0.001	0.01	0.003	-0.003
Min	-0.98	-0.5	-0.44	-0.63	-0.71
Max	8.67	8.2	7.30	7.4	8.3
Range	9.65	8.7	7.74	8.03	9.01
Std. Dev.	1.001	1.3	0.74	1.51	1.4

After the missing rate was increased to 60%, results as seen in Table 3 showed that CKR estimated data sets produced results with the best correlation estimate of 0.34. The coefficients of correlation produced by using Simple regression, mean and K-NN were 0.31, 0.3 and 0.29 respectively. CKR had the best mean estimate from the single imputation methods with a value of -0.12, values from KNN, Simple Regression and mean were the next in line with 0.14, 0.15 and 0.11 respectively. The proposed CKR produced a data set with a range of 8.51. K-NN, mean and simple linear regression produced data sets with ranges of 8.22, 8.7 and 7.81 respectively. Our proposed method alsp produced data points with a minimum of -0.12. Other single imputation techniques like K-NN, Mean and simple regression had there least figures as -0.23,-0.1 and -0.4 respectively. Simple linear regression, CKR and KNN methods produced maximum estimates of 7.99, 7.9 and 7.71.

Table 3. The Four Parameters Before and After Estimation from MCAR with Techniques at 60% rate

	MCAR @ 60%				
Parameter	Single Imputation Technique				
					Proposed Method
	Full Data	Simple Reg	Mean	K-NN	CKR
r	0.40	0.31	0.30	0.29	0.34
Mean	-0.0001	0.11	0.15	0.14	-0.12
Min	-0.98	-0.4	-0.1	-0.23	-0.61
Max	8.67	7.99	8.1	7.71	7.9
Range	9.65	8.5	7.81	8.22	8.51
Std. Dev.	1.001	1.28	0.66	1.39	1.3

3.1.2 Comparison of Parameter Estimates

The MAE and RMSE values are shown in tables 4, and 5. Small values are in general preferable as they imply better accuracy of missing data techniques. We earmarked (in boldface) small MAE values, with those less than or equal to (≤) 0.01 being indicative of methods with good precision.

Results of Table 4 show error rates from estimations of MCAR simulated datasets at 10 %, In general low MAE values were from CKR and K-NN techniques which were each 0.01. Simple regression and mean imputation techniques had MAE values of 0.02 and 0.04 respectively. Using the CKR method gave us an RMSE of 0.047. K-NN, Simple regression and the mean imputation gave us RMSE values of 0.048, 0.064 and 0.074 respectively.

Table 4. Errors between Original and Predicted Data from MCAR at 10% rate

	MCAR @ 10%			
	Single Imputation			
Error				*Proposed Method*
	Mean	Simple Reg.	K-NN	CKR
MAE	0.04	0.02	**0.01**	**0.01**
RMSE	0.074	0.064	0.048	0.047

MAE values in boldface are less than or equal to (≤) the 0.01 threshold

After Intensifying the missing-ness to 60% as seen in Table 5, The proposed CKR method gave us an MAE of 0.02 and an RMSE of 0.073. For the K-NN method MAE value was 0.07 while simple regression and mean imputation had values of 0.05 and 0.09 respectively. The RMSE value from using the K-NN method was 0.101. the mean imputation technique had the highest RMSE with a value of 0.117 and that for simple regression was a value of 0.078 which was higher than that of our proposed CKR method.

Table 5. Errors between Original and Predicted Data from MCAR at 60% rate

	MCAR @ 60%			
	Single Imputation			
Error				*Proposed Method*
	Mean	Simple Reg.	K-NN	CKR
MAE	0.09	0.05	0.07	0.02
RMSE	0.117	0.078	0.101	0.073

4.0 DISCUSSION

The missing mechanism considered in this study was MCAR at two (2) missing rates (low or 10% - high or 60%) which was simulated on a bivariate dataset with a univariate missing pattern on the outcome variable V2 after which the techniques were applied and data analysis on estimated data took place. The performance of the methods were compared regarding parameter estimates such as correlation coefficients, means, standard deviation/error, minimum, maximum and range alongside MAE and RMSE error metrics.

Results show that all single imputation techniques tended to produce consistent parameter estimates in MCAR simulated data sets at all missing rates considered which was expected since the methods didn't have to deal with problems of non-normality[19].While this is so, it is important to consider that the precision of all techniques reduced slightly as missing-ness increased from 10% to 60%. The mean imputation also produced reasonable estimates, which is largely due to the complete randomness of our missing values making the reduced sample a random subset of our original data as suggested by[20]. Our proposed CKR regression performed as the best among all methods considered as it gave closer estimates to the original and didn't grossly underestimate our standard error as presumed. MAE rates for the CKR and MICE technique fell on and below the postulated threshold of 0.01 respectively. The simple regression and K-NN techniques in general faired better than mean imputation which had the highest MAE and RMSE rates from 10% to 60 % missing-ness. The findings are consistent with those in reviewed literature and confirm their recommendations.[1], [14], [18], [21].

4.1 CONCLUSION

In accordance to our aim of developing and investigating the efficacy of CKR, it was found to be well suited for MCAR mechanism as it outperformed other single imputation techniques, this was evident in the nearness of its parameter estimates to that of the original data and its relatively low MAE and RMSE rates, the performance of K-NN and Simple regression were very nearly at par. The slight superiority of the CKR over the two previously mentioned techniques was attributed to the idea that the concept of CKR is mainly the amalgamation of them both with only nuances in execution the proposed CKR also proved to be the most robust among all single imputation techniques as changes in its error rates while increasing missing proportions where low. The CKR technique was concluded to be an effective single imputation technique in comparison to its counterparts considered in this study, it was seen to perform its very best in MCAR conditions having low missing rates of about 10%.

CONSENT (WHERE EVER APPLICABLE)

NOT APPLICABLE

ETHICAL APPROVAL

NOT APPLICABLE

REFERENCES

[1] Batista, G. E. A. P. A., & Monard, M. C. (2003). "An Analysis of Four Missing Data Treatment Methods for Supervised Learning." *Applied Artificial Intelligence* 17(5–6): 519–33.

.

[2] Kang, H. (2013). "The Prevention and Handling of the Missing Data." *Korean Journal of Anesthesiology* 64(5): 402-6.

.

[3] Vaishnav, R. L., & Patel, K.M. (2015). "Analysis of Various Techniques to Handling Missing Value in Dataset." *International Journal of Innovative and Emerging Research in Engineering* 2(2): 191–95.

[4] Nwakuya, M. T. & Nwabeueze, J. C. (2016). "Relative Efficiency of Estimates Based on Percentages of Missingness Using Three Imputation Numbers in Multiple Imputation Analysis." *European Journal of Physical & Agricultural Sciences* 4(1): 63–69.

[5] McKnight, P. E., McKnight, K. M., Sidani, S. & Figueredo, A. J. (2007). **Missing Data: A Gentle Introduction**. The Guilford Press. New York London.

[6] Buchman, S. (2018). "Overview of Approaches For Missing Data".

[7] Siddique, J., Harel, O., & C. M. Crespi.(2012). "Addressing Missing Data Mechanism Uncertainty Using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial." *Annals of Applied Statistics* 6(4): 1814–37.

[8] Little, R. J. A. & Rubin, D. B. (1997). **Statistical Analysis with Missing Data.** A John Wiley & Sons, Inc.,Publication.

[9] Schafer, J. L. (1997). **Analysis of Incomplete Multivariate Data**. Library of Congress Cataloging in Publication Data.

.

[10] Guan, N. C., & Yusoff, M. S. B., (2011). "Missing Values in Data Analysis: Ignore or Impute?" *Education in*

Medicine Journal 3(1): 6–11.

[11] Little, R. J. A. & Rubin, D. B. [Ed] (2002). **Statistical Analysis with Missing Data.** A John Wiley & Sons, Inc.,Publication.

[12] Paniagua, D., Amor, P. J., Echeburua E. & Abad, F. J. (2017). "Comparison of Methods for Dealing with Missing Values in the EPV-R." 29(3): 384–89.

[13] Grzymala-Busse, J, W. and Hu, M.(2001). "A Comparison of Several Approaches to Missing Attribute Values in Data Mining" *Rough sets and current trends in computing* 2005 (Chapter 46): 378–85.

[14] Truxillo, C. (2002). "A Comparison of Missing Data Handling Methods." *SAS® Institute Inc, Cary, NC*

[15] Acuña, E. &, Rodriguez, C (2004). "The Treatment of Missing Values and Its Effect on Classifier Accuracy." *Classification, Clustering, and Data Mining Applications* (1995): 639–47.

[16] Nirelli, L, M., Larsen, M. D., Croghan, I. T., Schroeder, D. R., Offord, K. P. & Hurt, R. D. (2005). "Comparison of Methods for Handling Missing Data in a Collegiate Survey of Tobacco Use." *Working Paper*: 3439–46.

[17] Schouten, R. M., Lugtig P., & Vink, G. (2018). "Generating Missing Values for Simulation Purposes : A Multivariate Amputation Procedure." *Journal of Statistical Computation & Simulation* 88(15): 2909–2930.

[18] Nookhong, J. and Kaewrattanapat, N.(2015). "Efficiency Comparison of Data Mining Techniques for Missing-Value Imputation." *Journal of Industrial and Intelligent Information* 3(4): 305–9.

[19] Nakai, M. (2011). "Simulation Study: Introduction of Imputation Methods for Missing Data in Longitudinal Analysis." *Applied Mathematical Sciences* 5(57): 2807–18.

[20] Nakai, M., Chen, D. C., Nishimura, K., & Miyamoto, Y. (2014). "Comparative Study of Four Methods in Missing Value Imputations under Missing Completely at Random Mechanism." https://pdfs.semanticscholar.org/620b/c6d3e15b2e5b252ef3b1f53c9148b6989148.pdf.

[21] Zhang, Z. (2016). "Missing Data Imputation: Focusing on Single Imputation." *Annals of translational medicine* 4(1): 9.