A Novel Cleansing Method for Random-Walk Data using Extended Multivariate Nonlinear Regression: A Data Preprocessor for Load Forecasting Mechanism

Hussein Bakiri
Institute of Finance Management (IFM)
Tanzania

Hamisi Ndyetabura
University of Dar es Salaam (UDSM)
Tanzania

Libe Massawe
University of Dar es Salaam (UDSM)
Tanzania

Hellen Maziku
University of Dar es Salaam (UDSM)
Tanzania

ABSTRACT
The efficiency of any load forecasting mechanism depends on the quality and distribution characteristics of the training data. Outliers and missing values are the primary concern, especially in developing countries' load data. Several research works have proposed the models for the imputation process to deal with outliers before forecasting. However, the efficiency of these approaches is compromised when it comes to data that falls into a random-walk distribution. Thus, this study aims to develop an efficient data cleansing model that accounts for a random-walk distributionby extending the Multivariate Nonlinear Regression (MNLR) method. The k-mean algorithm is used to detect and analyze the size of an outlier in the data. Twenty-minutes interval load data from 2015 to 2019 collected at Kinondoni-North (at Mikocheni distribution network in Dar es salaam) is used in this study. After analyzing the data for outliers, the empirical results detect the presence of outliers by 5.17852% (which is 5207 out of 105192). Finally, the extended-MNLR (e-MNLR) modelachieves promising results over the ANN, SVM, Miss Forest, MICE, and KNN algorithms by attaining 2.109137, 1.956039, and 7.787976 values of RMSE, MAE, and MAPE, respectively.

General Terms
Outliers, Data Cleansing

Keywords
Load forecasting, Developingcountries, Outliers, Data cleansing, extended-MNLR

1. INTRODUCTION
Load data in many developing countries is characterized by outliers and missing values caused by the inefficiency of transmission technologies, frequent power cuts and monitoring devices [1], [2]. The analysis conducted by [3]on 22developing countries indicatesthat;of all the proposed load forecasting models, 63.64% considered data cleansing preprocessors before prediction[3].Moreover, many data cleansing mechanisms have been proposed ranging from statistical (mean and median imputation[4]), regression (MNLR and MICE[5]–[7]) and machine learning (KNN[8], ANN, SVMand Miss-forest imputation[9]) models.However, those above statistical, regression and machine learning methodshave some limitations regarding the nature of data distribution. For example, the statistical imputation techniqueis not efficient if the missingness of the data is entirely at randomandwhen the distribution is not linear[4].Furthermore, the existing soft computing techniques do not efficiently handle dataset whose characteristics exhibits a random walk behavior[10]. Therefore, this study aims to establish a novel data cleansing mechanism that suites the random-walk behaviour exhibited in the Tanzania context using an e-MNLR.

The results of outlier analysis in this study have revealed abnormal data that might mislead decision-makers and load analysts if not taken care of. Moreover, outliers alert the utility company's decision makers and load analysts to pay attention to data preprocessing to get rid of unreliable results. Furthermore, the proposed e-MNLR model will benefit utility companies andthe entire data analysts to achieve a better data cleansing process.

2. LITERATURE REVIEW
2.1 Load Characteristics in Developing Countries
Researchers report the existence of peculiar long-term load distribution in developing countries contrary to the developed ones. Such abnormality may be caused by many factors, including social-economic and technological grounds[2], [11]–[13]. Theexhibited abnormal load distribution in developing countries (characterized by nonstationary and random-walk behaviors) hinders the efficiency of the existing forecasting methods. The authors[2], [14] report further that the presence of outliers and missing values might be caused by the intermittent power shortages,ageing of both transmission and data reading devices.

The characteristics of the electricity load in east African countries (Kenya and Tanzania) has been investigated by [15]. The authors in the former study observed that the load variationconcerningthe time of the day is nonstationary. Moreover, in this study, the long-term characteristic of load variation in Tanzania is investigated by testing the data for a randomwalk test. The test result indicates a random-walk behavior in the load data for 2015 to 2019, as presented in Fig 6.

2.2 Determinants of Electricity Consumption in Developing Countries
Investigation of factors affecting electricity consumption in developing countries has been studied in the research works by [2], [11], [16]–[20]. The findings from the works above indicate that GDP, population and electricity price are common factors affecting long-term electricity consumption. In addition to that, temperature, daytime and calendar events seem to be common factors affecting short-term consumption[3]. Moreover, further research findings on the load demand determinants can be found in work conducted by [3].

2.3 DataCleansing Techniques

[21] proposes a method for choosing neighbouring critical values when searching and replacing outliers using the Mahalanobis distance technique. [22] proposes an M-Estimators of Huber and Hampel algorithm, which finds the effective outlier replacer. A procedure for imputing values missing at random in a complex data structureis proposed by [23] using sequential regression multivariate imputation. [24]proposes a parimputation algorithm to improve the efficiency of the imputation process in case the neighbours are far from the missing data.[25]employs different imputation techniques to estimate replacer values and applies an evaluator approach to identify the permanent conclusion.[26] applies an unsupervised neural network algorithm called Kohonen Self-Organizing Maps (KSOM) replacing outliers in sludge wastewater treatment plants.

[27]proposes an approach for detecting, removing and fill the outliers in a time series data using Winsorising, Discrete Fourier Transform(DFT) and Inverse Fourier Transform (IFT). A method for dealing with missing values in heterogeneous data using k-Nearest Neighbors is introduced [28].[9] proposes a robust non-parametric method using the MissForest algorithm for imputing missing values based on different data types, unlike in other methods such as k-NN, which is limited to continuous data type. Lastly, [5] proposes general multiple imputations based on three phases: imputation, analysis, and pooling using Multivariate Imputation by Chained Equation (MICE).

2.4 The MNLR Model

The MNLR model has been proposed in the research work by [29]. The MNLR has produced reasonably good results when applied to nonlinear time series data. Moreover, the model can be extended through parameterization. The general form of the MNLR model is shown in equation (2).

$$Y = \alpha_0\left(x_1^{\alpha_1}\right)\left(x_2^{\alpha_2}\right)\ldots.x_n^{\alpha_n} \quad (2)$$

Where α_0 - α_n model parameters, x_1 - x_n are input variables, and Y is a dependent variable.

2.5 Parameter Estimation Techniques

Identifying values of model parameters is one of the challenging tasks in regression analysis. Different methods for parameter estimation, including rank regression (least squares), maximum likelihood estimation and Bayesian estimation, have been proposed in the literature[30]. Nonlinear Least Square (NLS) is a widely used parameter estimation method. The objective function of NLS is to minimize the sum of squared residuals (see equation (3)).

$$S = \sum_{i=1}^{m} {r_i}^2 \quad (3)$$

$$r_i = y_i - f(x_i - \beta) \quad (4)$$

Where r_i= residual in prediction error, β= model parameter, x_i= independent variable and y_i= dependent variable. Therefore the value of β can be estimated derivatively using equation (3) and equation (4)[30].

3. MATERIAL AND METHODS

3.1 Research Design

This study conducts a confirmatory analysis to test the threeclaims observedfrom works of literature; firstly, it is claimedthat load data in manydeveloping countries contain outliers. The second claim is that load data in many developing countries exhibit a nonstationary (random-walk) distribution trend. The third claim is that existing Outlier Imputation (OI) techniquesmay not yield efficient results fora situation where the two claims mentioned above are exhibited.

The existing outlier imputation methods are applied to the available data to investigate their forecast capability.After that, a model is extended repetitively using the heuristic technique until the prediction error is minimum. The overall research design process is shown in Fig 1.

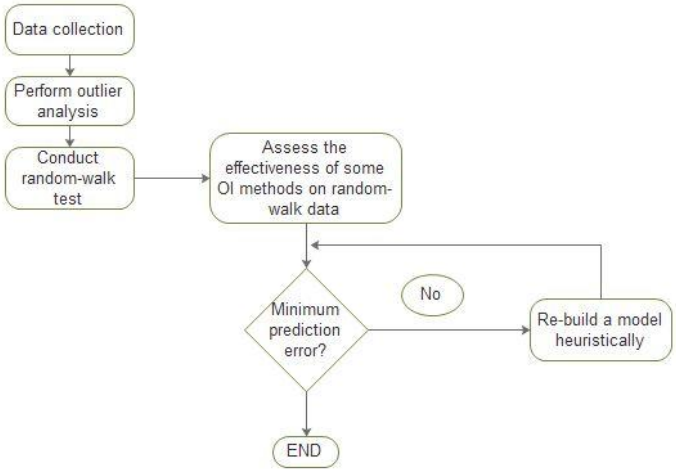

Fig 1:The flowchart for the research designprocess of the proposed outlier imputation mechanism

3.2 Data Collection

The twenty-minute interval load data from 2015 to 2019 has been collected from TANESCO. The data is acquired from the Accumulated Meter Reading (AMR)situated at the three transformers (BBQ Village-SS2, Kimweri Avenue-SS2 and Abiudi Street-SS2). The annual GDP data from 2015 to 2019 is collected from the National Bureau of Statistics (NBS). The monthly number of customers and population data from 2015 to 2019 is also collected from TANESCO. The three-hour temperature data (from 2015 to 2018) is collected from Tanzania Meteorological Authority (TMA).

3.3 Outlier Detection and Analysis

The k-mean algorithm is used to examine the presence of outliers in the electricity load data. The choice of the k-mean algorithm is based on the fact that the number of clusters and centroid values are known in advance from expert knowledge. Furthermore, after conducting load analysis using the five-year data, the average minimum consumption is observed to be 16kWh, from which the definition of outlier is based.Thus, two clusters are formed for which two centroid values (4 and 25) are initialized.

3.4 The Random-Walk Test

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are used to test for random-walk behaviour in the time series data. Furthermore,a load graph against time using the five years'data (2015 to 2019) is also plotted to aid the data analysis process.

3.5 Determining the Capability of the Existing Forecasting Methods on Tanzania Data

The design of the model has been established by applying existing statistical (mean and median imputation), regression-based imputation (MICE and MNLR) and machine learning algorithms (ANN, SVR, KNN, Miss Forest) methods to the available data. The results are then tested and validated to identify model errors. If the error is reasonably significant in all the chosen methods, then either of the models is extended to obtain promising results. The easilyextensible model is taken into consideration for the adaptation process. Furthermore, the process of selecting the model to extend is based on the existing literature concerning the nature of the datacharacterized by random-walk behaviour.

4. MODELING THE PROPOSED e-MNLR METHOD

4.1 Design Requirements of the Proposed e-MNLR Model

4.1.1 Requirements as per works of literature

GDP, population, and the number of customers have been found to exert a significant long-term effect on power consumption in developing countries as per the research study conducted by [3]. Moreover, in a similar study, the authors identified temperature, day type, and daytime as the primary short-term drivers of electricity consumption in developing countries. Furthermore, handling the outliers and missing values is crucial for designing load forecasting mechanisms in developing countries, as reported in the study [2].

4.1.2 Confirmatory Analysis on Load Characteristics in Tanzania

A. Determinants of Electricity Consumption in Tanzania

The determinants of electricity consumption in Tanzania are identified by plotting line graphs of load versus GDP, customers, and population. From the graph plot in Fig 4, Fig 5, and Fig 7, it is clear that GDP, number of customers and population are the main factors influencing electricity consumption in Tanzania. Moreover, the temperature seems not to correlate with electricity consumption when it comes to the residential buildings, as reported in the research work by [3]. In addition to that, the effect of daytime on short-term consumption is reported to be significant in a similar study by [3]. Furthermore, theshort-term effect of day type on electricity consumption is shown by the graph plot in Fig 6.

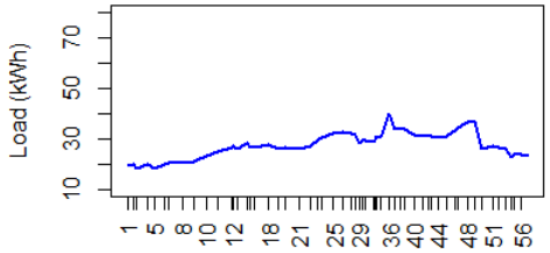

Fig4:Monthly load vs number of customers from 2015-2019 (56 months)

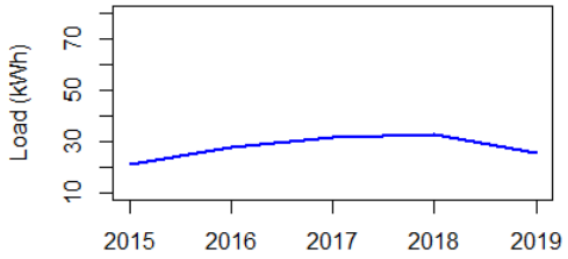

Fig5: Yearly load vs GDP

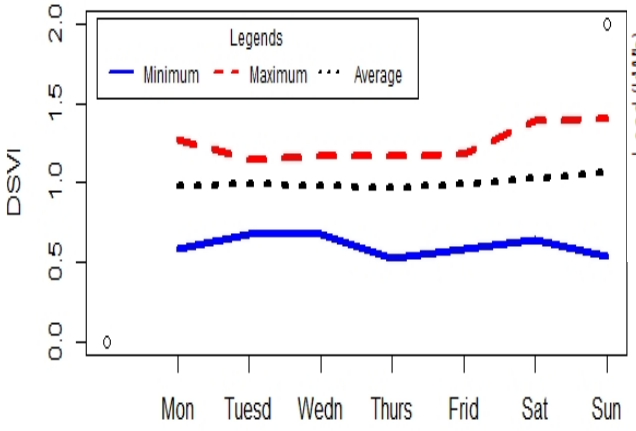

Fig6: The day type effect on electricity consumption as per the study area in Dar es Salaam Tanzania from 2015 to 2019

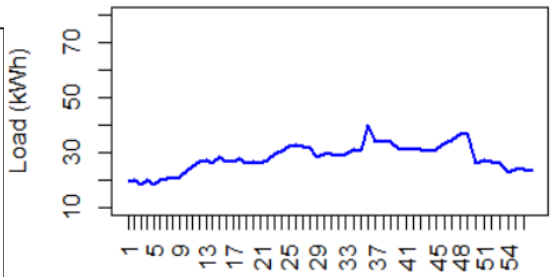

Fig 7: Monthly load vs area's population from 2015-2019 (54 months)

B. Characteristic Load Distribution in Tanzania

The test conducted on 2015 to 2019 data indicates that the load distribution trend in Tanzania is characterized by random-walk behaviour. Firstly, from Fig 8-a (autocorrelation versus lags), it can be seen that a gentle decline for the number of lags is observed that span toward the bottom margin of a significant correlation line. Secondly, from Fig 8-b (partial autocorrelation versus lags), the first lag shows a significant correlation while the rest are insignificant. Furthermore, Fig 8-c confirms the nonstationary behaviour exhibited by load data. Therefore, the graph mentioned above analysis confirms the random-walk property in the load data.

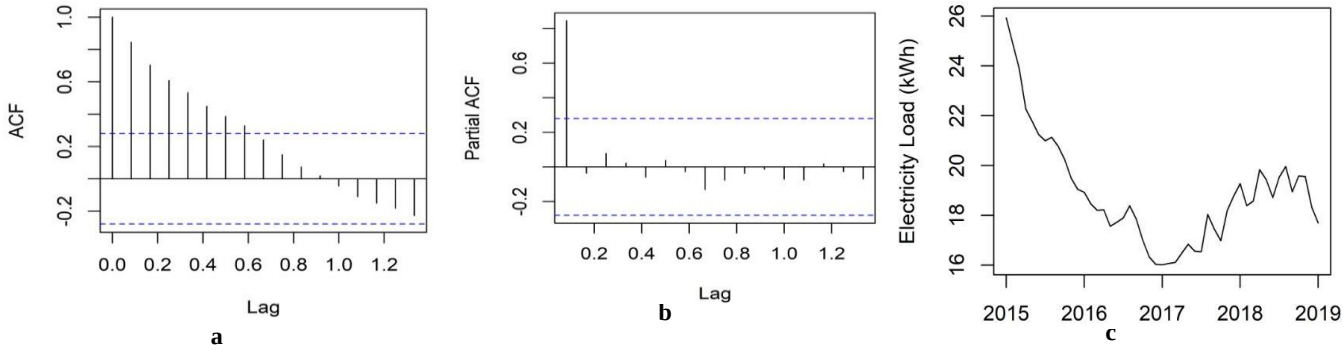

Fig8: Random-walk test results using 2015 to 2018 North-Kinondoni load data in Dar es Salaam

C. Outlier Analysis in Tanzania Load Data

After running the k-mean algorithm in R, empirical results show that; on 105192 data, the load was found to contain 5.17852% (which is 5207 out of 105192) outliers. This finding conforms with currentresearch works that load data in developing countries are most likely to contain corrupt entries.

D. Analyzingthe Capability of Existing Methods on Tanzania Data

Eight data cleansing methods were executed on twenty-minutes interval Tanzania load data using the R language, and results are shown in Table 1. From the table, it is evident that the results are not promising.

Table 1:Execution environment for the eight widely used imputation methods during the comparative evaluation process

Cleansing Method	**Execution Environment**	**RMSE**	**MAE**	**MAPE**
Mean	N/A	8.404473	6.956038	24.06969
Median	N/A	9.131619	6.772813	20.63348
MNLR	N/A	12.77399	9.226775	34.34231
Miss-Forest	8GB RAM, Co-i5, Max. iterations = Default, Sample Size = 16055, Training Data = 84154	3.035463	2.732547	10.24622
SVM	8GB RAM, Co-i5, Kernel = radio, Type = eps-regression, Training Data = 5000	5.32299	3.15148	11.35937
ANN	8GB RAM, Co-i5, No. Hidden Layers = 2, Hidden Neurons = 4, Threshold = 0.001, Algorithm = rprop+, Activation Function = Logistic, err.fct = sse, StartWeight= 1.3, StepMax = 20,000	6.9456	7.466061	27.8525
MICE	8GB RAM, Co-i5, Max. iterations = 150, M = 5, Method = sample	12.09854	9.361562	33.43858
KNN	8GB RAM, Co-i5, variable = load, k = 6	2.960205	4.006625	14.47315

4.2 Re-building the e-MNLR Model

From the design requirements outlined in the previous section, the architectural model for the proposed data cleansing is created as presented in Fig 9.

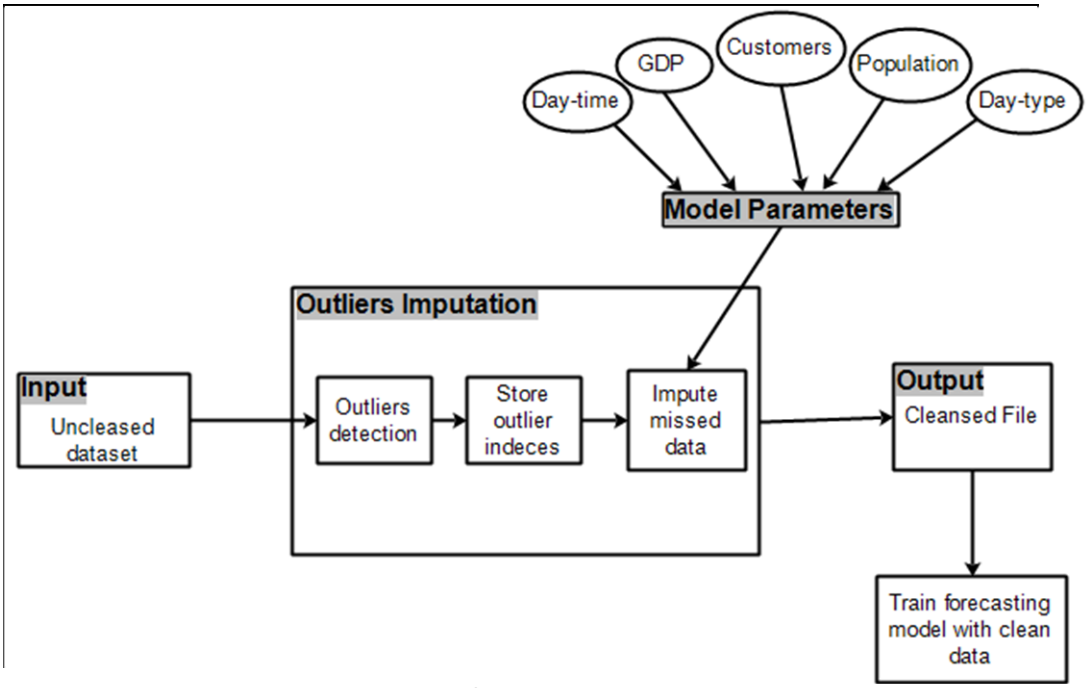

Fig 9:**The architectural model of the proposed data cleansing mechanism.**

4.2.1 Calibrating the Daytime Parameter

Concerning the study conducted by [3], the daily electricity consumption in Tanzania varies accordingly with the time of the day. Their findings indicate that there is high consumption during night hours, and it becomes lower during midnights.In addition to that, it decreases gradually from night into dawn, and there is an abrupt increase from the early morning. This fact conveys the characteristic nature of electricity consumption withthe time of acorresponding day.

The MNLR in equation (2) is used to estimate the optimal time values. The equation is parameterized to include GDP, population, number of customers, time of the day and day type,as shown in equation (5). Using the NLS method presented in equation (3) and equation (4), the model parameters of equation (5) can be estimated. Furthermore, using equation (6), the optimal daytime values can be calibrated.

$$y_t = x_t^{\alpha_t} * x_d^{\alpha_d} * x_g^{\alpha_g} * x_p^{\alpha_p} * x_c^{\alpha_c} \quad (5)$$

$$x_t = log^{-1}\left(\frac{([log(y_t)] - [\alpha_d * log(x_d) + \alpha_g * log(x_g) + \alpha_p * log(x_p) + \alpha_c * log(x_c)])}{\alpha_t}\right) \quad (6)$$

Where y_t = load at a time (20 minutes), x_t = time value whose load is sought, x_d = day category number, x_g = GDP value of the current year, x_p = current population,x_c = current number of customers., and $\alpha_t, \alpha_d, \alpha_g, \alpha_p and \alpha_c$ are time, day, GDP, population and customer parameters, respectively.

After analyzing time variation and deducing the difference of the time values, the stepwise function shown in equation (7) is then established to generalize the time values of each twenty-minute interval for the entire day. The resulting time values for the entire day is shown in Table 2.

$$f(x_{cat}) = \begin{cases} x_{mid} - 0.1, & x_{cat} = midnight \\ x_{lmid} - 0.05, & x_{cat} = late\ midnight \\ x_{dawn} - 0.05, & x_{cat} = dawn \\ x_{morn} - 0.025, & x_{cat} = morning \\ x_{lmorn} + 0.1, & x_{cat} = late\ morning \\ x_{noon} - 0.05, & x_{cat} = noon \\ x_{afnoon} - 0.05, & x_{cat} = afternoon \\ x_{evening} - 0.5, & x_{cat} = evening \\ x_{levening1} + 0.15, & x_{cat} = late\ evening1 \\ x_{levening2} + 0.15, & x_{cat} = late\ evening2 \\ x_{night} - 0.1, & x_{cat} = night \end{cases} \quad (7)$$

Table 2: The optimal time values after the time-transformation process

Time category	Time interval	Optimal time value	Difference
Midnight	0.00 - 2.00	4.2	-0.1
		4.1	
		4	
Late Midnight	2.20 - 3.40	3.55	-0.05
		3.5	
		3.15	
Dawn	4.00 - 5.40	3.3	-0.025
		3.1	
		2.975	
Morning	6.00 – 9.20	3.075	-0.025
		3.5	
		3.025	
Late Morning	9.40 – 11.40	2.8	0.1
		2.875	
		2.975	
Noon	12.00 – 13.40	3.475	-0.05
		3.425	
		3.375	
Afternoon	14.00 – 15.40	3.175	-0.05

		3.125	
		3.075	
Evening	16.00 – 17.20	3.25	-0.05
		3.2	
		3.15	
Late Evening1	17.40 – 19.00	3.2	0.15
		3.35	
		3.5	
Late Evening2	19.20 – 21.00	3.95	0.15
		4.1	
		4.25	
Night	21.20 – 23.40	5.1	-0.1
		5	
		4.9	

4.2.2 Consideration of the Day Type

From the graph plot of daily load variation presented in Fig 6, the electric load variesby day type. From the graph, the electricity consumption on Saturday is considerably higher than that on weekdays, while that on Saturday seems to be slightly lower than that on Sunday. Moreover, using the observations above, the consumption is categorized into weekdays (Monday to Friday), weekend1 (Saturday), and weekend2 (Sunday). Therefore, weekday, weekend1, and weekend2 are given the ordinal values of 1, 2, and 3, respectively.

4.2.3 Thee-MNLRModel

The calibrated time values, GDP, number of customers, population and day type are then used to re-estimate the optimal parameter values of the model. Therefore,the new extended version of equation (5) is presented by equation (8).

$$y_t = x_t^{0.2271} * x_d^{0.0456} * x_g^{-0.4465} * x_p^{-1.7216} * x_c^{3.7746} (8)$$

Where, y_t = Load at 20-minutes interval at the year n, x_t= Time of day*n*, x_d = Day of a month, x_g = real-time GDP value,x_p= real-time population, and x_c= real-time number of customers. The model is then implemented using R programming environments.

5. RESULT AND DISCUSSION

5.1 Results

The simulation results after the implementation of the e-MNLR model are compared with two statistical(mean imputation and median imputation), two regressions (MICE and MNLR), and four machine learning (KNN, ANN, MissForest, and SVM) models. Twenty cases from each month (January to August 2019) are picked randomly from a sample of 160 records during the model evaluation process.

Table 3 shows evaluation results when the e-MNLRwas compared to the rest of the chosen methods. As presented in the table, it can be observed that the e-MNLR model has the lowest RMSE, MAE, and MAPE values compared to its counterparts. The MAPE value (7.8%) indicates better accuracy of the developed e-MNLR model; MissForest (10.2%), KNN (14.5%), and SVM (11.4%) as also presented by the bar chart in Fig 10.

Table 3: Evaluation results based on eight imputation methods

Imputation Method	RMSE	MAE	MAPE
Mean	8.404473	6.956038	24.06969
Median	9.131619	6.772813	20.63348
MNLR	12.77399	9.226775	34.34231
Miss-Forest	3.035463	2.732547	10.24622
SVM	5.32299	3.15148	11.35937
ANN	6.9456	7.466061	27.8525
MICE	12.09854	9.361562	33.43858
KNN	2.960205	4.006625	14.47315
e-MNLR	2.109137	1.956039	7.787976

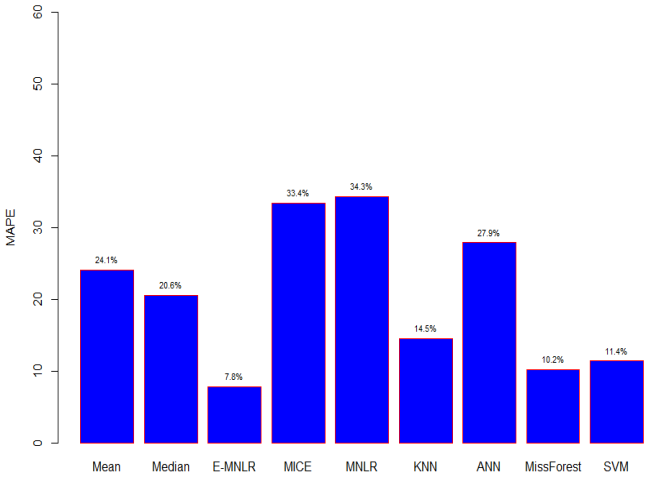

Fig10:The bar chart of MAPE values against imputation methods

The performance of the developed e-MNLR model relative to mean and median imputation can also be portrayed through visual analysis. From the line graph in Fig 11, the graph on the left-hand side indicates the superiority of the e-MNLR method over the mean imputation method. The same case is observed from the graphon the right-hand side, where the e-MNLR performance is compared to the median imputation method.

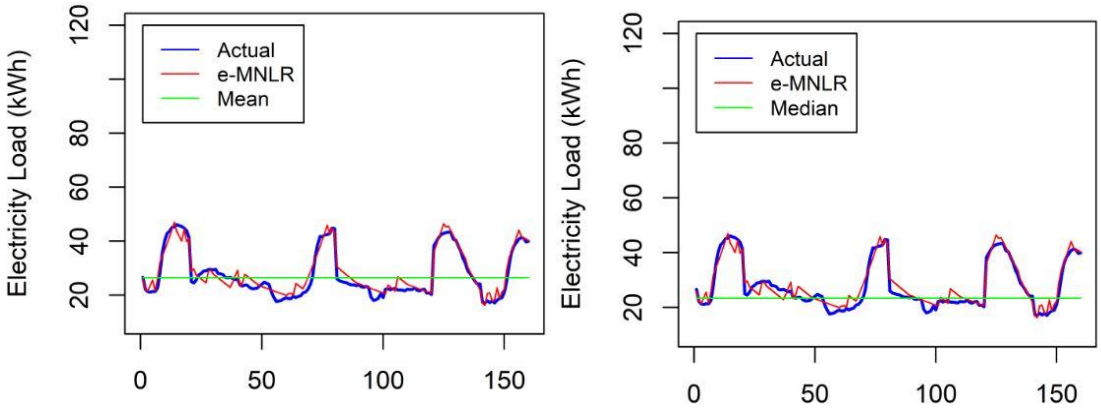

Fig 11:**Evaluation results of statistical versus e-MNLR on 160 records (January to August 2019)**

The performance of the developed e-MNLR model to regression methods is presented using graph plots in Fig 12. As presented in Fig, the graph on the left-hand side indicates the superiority of the e-MNLR method over the MNLR method. The same case is observed from the graph on the right-hand side, where the e-MNLR performance is compared to the MNLR method.

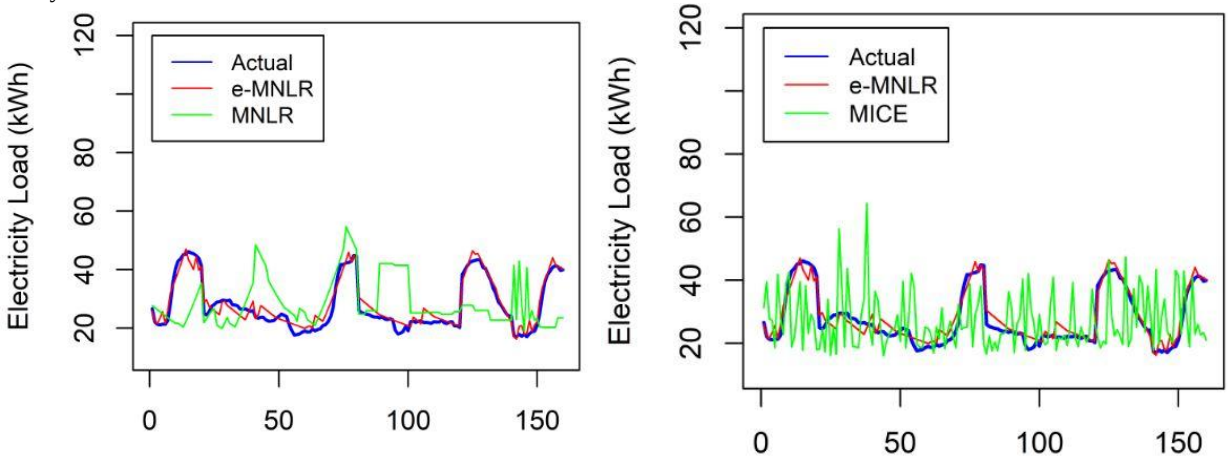

Fig12:**Evaluation results of regression models versus e-MNLR on 160 records (January to August 2019)**

The performance of the developed e-MNLR relative to machine learning models is presented using graph plots in Fig 13 and Fig 14. As presented in Fig 13, the graph on the left-hand side indicates the superiority of the e-MNLR method over the ANN model. The same case is observed from the graph on the right-hand side, where the e-MNLR performance is compared to the SVM model. Furthermore, the performance of the e-MNLR model is also compared to KNN and MissForest algorithms, as presented by graphs in Fig 14.

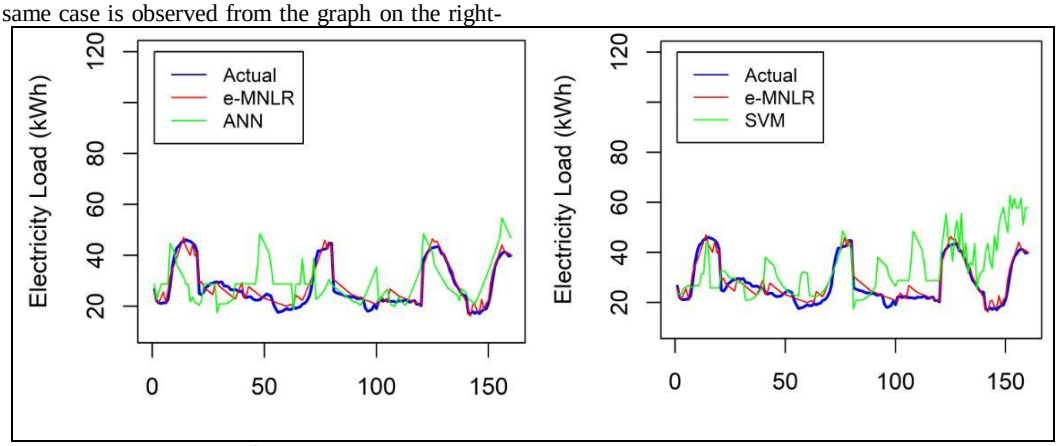

Fig 13:**Evaluation results of ANN and SVM models versus e-MNLR on 160 records (January to August 2019)**

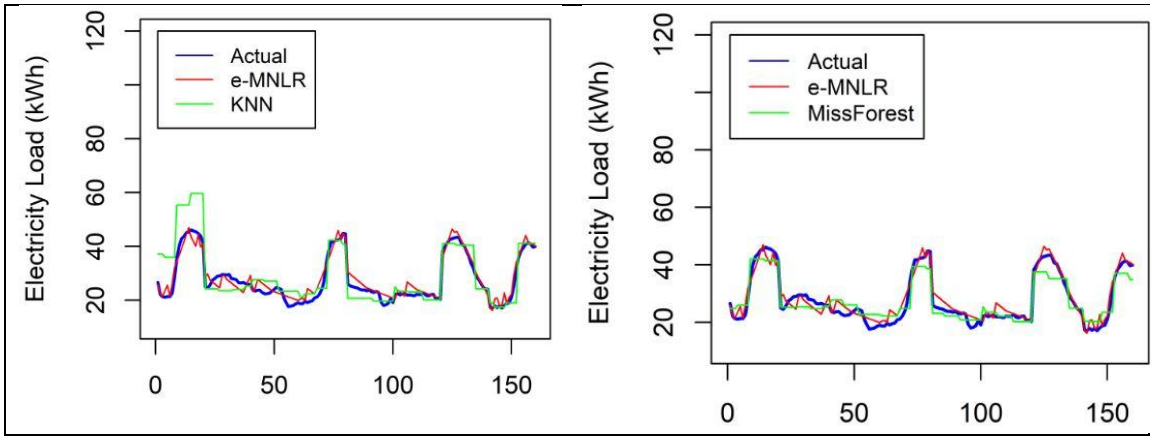

Fig14.Evaluation results of KNN and MissForest models versus e-MNLR on 160 records (January to August 2019)

Another interesting phenomenon exhibited by the e-MNLRmodel contrary to its counterparts is that the maximum error is more minor throughout the validation data (see Table 4). Having obtained the minimum value of maximum absolute error implies that the model has not deviated much from the average error. This intensifies the strength of the developed model over the rest of the methods.

Table 4: Maximum and Minimum absolute errors for all nine methods

Method	Max. AE	Min. AE
Mean	19.69733	0.002671
Median	22.81	0.01
MNLR	25.05	0.03624
Miss-Forest	9.455689	0.017568
SVM	36.42	0.786627
ANN	25.9473	0.15359
MICE	38.17	0.0
KNN	19.81	0.005
e-MNLR	5.382587	0.024846

5.2 Discussion

This research has identified the presence of random walk behaviour exhibited by long-term load profiles in Tanzania. This behaviour is mainly caused bythe undetermined annual increase in consumption demand brought by an unstable economy and outliers in the data. Furthermore, this work has revealed that the random walkbehaviourlimits the efficiency of existing data cleansing methods such as statistical, regression and machine learning. The lower values of RMSE (2.109), MAE (1.956), and MAPE (7.787) presented in Table 3 implies higher accuracy of the developed e-MNLR cleansing model to the counterparts. In addition to that, the small number of maximum absolute error (5.382) presented in Table 3 indicates how graceful the model deviates from the average actual values, contrary to counterparts. Thus, the method presented in this paper has outperformed other models such as statistical, regression, and machine learning algorithms.

The model in this paper has exhibited peculiar capability of predicting the missing values and outliers in the situation where data is characterized by random walk behaviour, where the existing methods could not. In addition to that,this study has focused on developing the data cleansing model,mainly based on residential buildings data due to the nature of the study area. Furthermore, due to the challenge of data availability, this research is confined to GDP, population, and the number of customer data as major economic indicators included in the model. For instance, the available electricity price and the number of customer data were just that of few years in such a way that its causality effect on electricity consumption could not be deduced effectively.

6. CONCLUSION

The efficiency of any prediction method depends on the nature of the data itself (stationary vs nonstationary, linear vs nonlinear). Most of the existing prediction methods tend to yield inappropriate results regarding nonstationary time series data that exhibit random walk behaviour. Thus, this work has tested eight imputation methods: ANN, mean, median, KNN, SVM, MICE, MissForest, and MNLR on Tanzania's load data that exhibits the random walk distribution properties. The test resulthasindicated that the existing methods could not predict effectively. Contrary, the developed e-MNLR model has yielded promising results and was found to be superior in the situation where load data is attributed by random walk behaviour compared to mean, median, ANN, MissForest, KNN, MICE, SVM, and MNLR methods.

Future research works may extend the output of this work to accommodate commercial areas and consider both the number of customers and electricity price factors.In addition to that, the proposed data cleansing model can be embedded in any forecasting mechanism to clean load data before prediction.

7. ACKNOWLEDGMENT

The authors would like to extend their sincere gratitude to Tanzania Electricity Supply Company (TANESCO) for the immense support through workshops and data provision. Secondly, intimate appreciation to Sida for masterminding smart grid researches in Tanzania. Lastly, it will be amess not to mention the entire iGrid research group from the University of Dar es Salaam for their willingness towards focus group discussions.

8. REFERENCES

[1] S. Saab, E. Badr, and G. Nasr, "Univariate modeling and forecasting of energy consumption: The case of electricity in Lebanon," *Energy*, vol. 26, no. 1, pp. 1–14, 2001, doi: 10.1016/S0360-5442(00)00049-9.

[2] M. U. Fahad and N. Arbab, "Factor Affecting Short Term Load Forecasting," *J. Clean Energy Technol.*, vol. 2, no. 4, pp. 305–309, 2014, doi: 10.7763/jocet.2014.v2.145.

[3] H. Bakiri, H. Maziku, N. Mvungi, N. Hamisi, and M. Libe, "Towards the Establishment of Robust Load Forecasting Mechanism in Tanzania Grid : Effect of Air Temperature and Daytime on Electricity Consumption in Residential

Buildings," *Int. J. Smart Grid*, vol. 5, no. 1, pp. 24–36, 2021.

[4] J. Sim, J. S. Lee, and O. Kwon, "Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications," *Math. Probl. Eng.*, vol. 2015, 2015, doi: 10.1155/2015/538613.

[5] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.

[6] J. Adamowski, H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada," *Water Resour. Res.*, vol. 48, no. 1, pp. 1–14, 2012, doi: 10.1029/2010WR009945.

[7] A. Yasar, M. Bilgili, and E. Simsek, "Water Demand Forecasting Based on Stepwise Multiple Nonlinear Regression Analysis," *Arab. J. Sci. Eng.*, vol. 37, no. 8, pp. 2333–2341, 2012, doi: 10.1007/s13369-012-0309-z.

[8] T. T. Dang, H. Y. T. Ngan, and W. Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," *Int. Conf. Digit. Signal Process. DSP*, vol. 2015-Septe, no. May 2016, pp. 507–510, 2015, doi: 10.1109/ICDSP.2015.7251924.

[9] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012, doi: 10.1093/bioinformatics/btr597.

[10] H. P. Sajjad, A. Docherty, and Y. Tyshetskiy, "Efficient Representation Learning Using Random Walks for Dynamic Graphs," 2019, [Online]. Available: http://arxiv.org/abs/1901.01346.

[11] S. C. Bhattacharyya and G. R. Timilsina, "Modelling energy demand of developing countries: Are the specific features adequately captured?," *Energy Policy*, vol. 38, no. 4, pp. 1979–1990, 2010, doi: 10.1016/j.enpol.2009.11.079.

[12] J. Steinbuks, "Assessing the accuracy of electricity production forecasts in developing countries," *Int. J. Forecast.*, vol. 35, no. 3, pp. 1175–1185, 2019, doi: 10.1016/j.ijforecast.2019.04.009.

[13] K. Kavaklioglu, "Modeling and prediction of Turkey's electricity consumption using Support Vector Regression," *Appl. Energy*, vol. 88, no. 1, pp. 368–375, 2011, doi: 10.1016/j.apenergy.2010.07.021.

[14] L. K. Hotta, "Effect of Outliers on Forecasting Temporally Aggregated Flow Variables," *Soc. Estad~stica e I~vestigacidn Oper.*, vol. 13, no. 2, pp. 371–402, 2004.

[15] N. J. Williams, P. Jaramillo, B. Cornell, I. Lyons-Galante, and E. Wynn, "Load characteristics of East African microgrids," *Proc. - 2017 IEEE PES-IAS PowerAfrica Conf. Harnessing Energy, Inf. Commun. Technol. Afford. Electrif. Africa, PowerAfrica 2017*, pp. 236–241, 2017, doi: 10.1109/PowerAfrica.2017.7991230.

[16] N. M. Odhiambo, "Energy consumption and economic growth nexus in Tanzania: An ARDL bounds testing approach," *Energy Policy*, vol. 37, no. 2, pp. 617–622, 2009, doi: 10.1016/j.enpol.2008.09.077.

[17] F. Egelioglu, A. A. Mohamad, and H. Guven, "Economic variables and electricity consumption in Northern Cyprus," *Energy*, vol. 26, no. 4, pp. 355–362, 2001, doi: 10.1016/S0360-5442(01)00008-1.

[18] G. Okoboi and J. Mawejje, "Electricity peak demand in Uganda: insights and foresight," *Energy. Sustain. Soc.*, vol. 6, no. 1, 2016, doi: 10.1186/s13705-016-0094-8.

[19] A. A. Aziz, N. H. Nik Mustapha, and R. Ismail, "Factors affecting energy demand in developing countries: A dynamic panel analysis," *Int. J. Energy Econ. Policy*, vol. 3, no. SPECIAL ISSUE, pp. 1–6, 2013.

[20] M. Khanna and N. D. Rao, "Supply and Demand of Electricity in the Developing World," *Annu. Rev. Resour. Econ.*, vol. 1, no. 1, pp. 567–596, 2009, doi: 10.1146/annurev.resource.050708.144230.

[21] K. I. Penny, "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance," *Appl. Stat.*, vol. 45, no. 1, p. 73, 1996, doi: 10.2307/2986224.

[22] K. J. Tvarlapati and K. A. Hoo, "A METHOD OF ROBUST MULTIVARIATE OUTLIER," *IFAC Proc. Vol.*, vol. 33, no. 10, pp. 641–646, 2000, doi: 10.1016/S1474-6670(17)38613-5.

[23] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Stat. Canada*, vol. 27, no. 1, pp. 85–95, 2001.

[24] S. Zhang and S. Member, "Parimputation : From Imputation and Null-Imputation to Partially Imputation," *IEEE Intell. Informatics Bullein*, vol. 9, no. 1, pp. 32–38, 2008.

[25] J. Ma, G. Zhang, J. Lu, and D. Ruan, "Impute missing assessments by opinion clustering in multi-criteria group decision making problems," *2009 Int. Fuzzy Syst. Assoc. World Congr. 2009 Eur. Soc. Fuzzy Log. Technol. Conf. IFSA-EUSFLAT 2009 - Proc.*, pp. 555–560, 2009.

[26] R. Rustum and A. Adeloye, "Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organizing Map," *J. Environ. Eng.*, vol. 133, no. 9, pp. 909–916, 2007, doi: 10.1061/(ASCE)0733-9372(2007)133:9(909).

[27] L. Plazas-nossa and A. Torres, "Detection of outliers and replacement of missing values in absorbance and discharge time series," in *10th International Urban Drainage Modelling Conference*, 2015, pp. 113–117.

[28] D. E. N. Frossard, I. O. Nunes, and R. A. Krohling, "An approach to dealing with missing values in heterogeneous data using k-nearest neighbors," 2016, [Online]. Available: http://arxiv.org/abs/1608.04037.

[29] G. Özbayoğlu and M. Evren Özbayoğlu, "A new approach for the prediction of ash fusion temperatures: A case study using Turkish lignites," *Fuel*, vol. 85, no. 4, pp. 545–552, 2006, doi: 10.1016/j.fuel.2004.12.020.

[30] P. J. Teusnissen and G, "Nonlinear least squares," *Manuscripta Geod.*, vol. 15, no. 3, pp. 137–150, 1990.