A comparison of Methods for Missing data treatment in building sensor data

Mehdi Pazhoohesh
School of Engineering
Newcastle University
Newcastle, UK
Mehdi.pazhoohesh@ncl.ac.uk

Zoya Pourmirza
School of Engineering
Newcastle University
Newcastle, UK
Zoya.Pourmirza@newcastle.ac.uk

Sara Walker
School of Engineering
Newcastle University
Newcastle, UK
Sara.walker@ncl.ac.uk

Data collection is a fundamental component in the study of energy and buildings. Errors and inconsistencies in the data collected from test environment can negatively influence the energy consumption modelling of a building and other control and management applications. This paper addresses the gap in the current study of missing data treatment. It presents a comparative study of eight methods for imputing missing values in building sensor data. The data set used in this study, are real data collected from our test bed, which is a living lab in the Newcastle University. When the data imputation process is completed, we used Mean Absolute Error, and Root Mean Squared Error methods to evaluate the difference between the imputed values and real values. In order to achieve more accurate and robust results, this process has been repeated 1000, and the average of 1000 simulation is demonstrated in this paper. Finally, it is concluded that it is necessary to identify the percentage of missing data before selecting the proper imputation method, in order to achieve the best result.

Keywords-energy and building data, data imputation; missing value; KNN; MCMC; MAE; RMSE.

I. INTRODUCTION

Nowadays, data collection is a key process in the study of Energy and buildings. For instance, Building Energy controls and retrofit analysis are two applications of collecting large amount of data from installed sensors. In addition, data collected from building has been used for modelling the energy consumption in buildings through different software such as EnergyPlus [1].

However, significant discrepancies between simulated and measured energy consumption of buildings is the motivation to focus more on analysing data collected through extensive sensor networks.

A. Related Work and Gap Analysis

Different calibration techniques such as Bayesian calibration [2], [3] and systematic evidence-based approaches [4] has been used to uncover discrepancies between simulated and measured energy consumption of buildings. However, a considerable amount of data are usually missed due to different reasons such as low signal-to-noise ratio, measurement error, malfunctioning of sensors, power outages at the sensors or network failure which can lead to data analysis problems. Hence, the estimating of missing values play a significant role in calibration of building energy models as a pre-processing step. Moreover, evaluation and prediction of building's energy consumption through statistical and data mining methods require time-series data in which missing values can significantly influence the analysis results, further emphasizing the importance of missing value estimation. Different approaches are used to deal with missing values in most scientific research domains such as Biology [5], Medicine [6] or Climatic Science [7]. However, there are limited studies to deal with missing data for the building energy system. One approach is to delete all missing values and analyse the behaviour of the building based on available data. The issue which may arise with this method is that there may be very few observations and a very small dataset to model the behaviour of the building based on that [8] [9]. Another approach is mean imputation, where missing data will be replaced with the mean value of all variables [8] [10]. This method distorts the distribution of the variable and also relationships between variables and can result in large errors between predicted and actual values. The other method used to treat the missing data is replacing missing data values with some constant (eg. zero). This has been used for the applications where they cannot tolerate having gaps between data [5]. Although ,a variety of techniques have been developed to treat missing values with statistical prediction in other fields, there is a lack of research concerning the substituting of missing values in order to provide guidelines to make the more appropriate methodological choice in energy and building related data. In the this study, we compare eight different imputation methods, namely, Monto Carlo Markov chain (MCMC) [11], Hmisc aregImpute [12], K-nearest neighbours (KNN) [13], simple Mean, Expectation-Maximization [14] [15], Random value, Regression and stochastic regression [15]methods, to find which method is the best fit for energy and building data sets. Comparison was performed on real lightning dataset collected from a 6 months period, under an Missing Completely at Random (MCAR) assumption and based on Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) evaluation criteria for estimating missing values in building data.

II. Method

A. Study Site

The data used for this study is lighting time-series data as the main dataset and corresponding occupancy data as the supportive dataset which were collected from the 3rd floor of Urban Science Building, Newcastle University, United Kingdom (Figure 1). Data collection took place between February 2018 and July 2018 at 1 minute intervals. The collected data were averaged to obtain half-hourly values with 7968 data points.

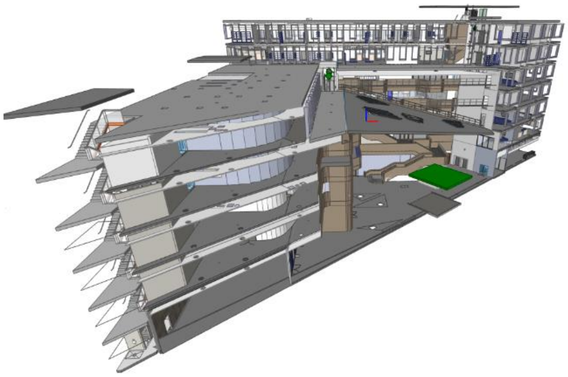


Figure 1: USB Building

B. Selection of Imputation Method

In order to conduct this study, we have selected eight imputation methods, which are the most well-known techniques that covers various statistical strategies in terms of simplicity to multiple imputation methods. These techniques are Mean, Random, Nearest Neighbour algorithm (KNN), aregImpute (Hmisc) in R, Markov Chain Monte Carlo (MCMC) [15], expectation-maximization (EM) algorithm [11], Regression and Stochastic regression methods. Here we briefly discuss each techniques. Mean method is based on imputation by replacing the missing data by the mean of all known values of that variable.

Random technique is used based on randomly predicting the missing values according to the maximum and minimum values of the dataset.

The nearest neighbour algorithm [16] is a nonparametric method which is used to replace the missing data for the variable by averaging non-missing values of its neighbours. In this method, K-nearest Neighbours are selected to predict the missing value and the influence is the same for each of these neighbours. Depends on the number of selected neighbours (K value), the estimated value could be significantly tolerated. Hence, choosing the proper number of neighbours, has great influence on the prediction. In this paper, the effect of different values of the parameter k on estimation accuracy is discussed.

The aregImpute function in the HMisc library [12]consists of replacing the missing value with predictive mean matching which is computed by optional weighted probability sampling from similar cases. In aregImpute function, missing values for any parameter are estimated based on other parameters. In this paper, occupancy data is considered as the supportive value for estimating the missing value in lighting dataset.

Markov Chain Monte Carlo (MCMC) is an iterative algorithm based on chained equations that uses an imputation model specified separately for each variable and involving the other variables as predictors. Monte Carlo Markov chain (MCMC) method is used to generate pseudo-random draws and provides several imputed data sets. MCMC requires either MAR or MCAR data sets and can be implemented on both arbitrary and monotone patterns of missing data. A Markov Chain is a sequence of possible variables in which the probability of each element depends only on the value of the previous one.

In MCMC simulation, by constructing a Markov chain that has the stationary distribution which is the distribution of interest, one can obtain a sample of the desired distribution by repeatedly simulating steps of the chain. Refer to Schafer [17] for a detailed discussion of this method.

In the regression imputation method, the missing values will be replaced with predicted score from regression equation. Although, the imputed data are computed using information from the observed data, only one representative value will be considered for each group of missing data which may result in weakens variance. Another method which is inspired from regression concept is stochastic regression method. This method aims to reduce the bias using additional step of augmenting each predicted score with a residual term.

Therefore, each missing value has a different imputed number to be replaced with [15].

III. Discussion

A. Evaluation Criteria

To evaluate the forecast, mean absolute error (MAE), and root mean square error (RMSE) were computed over the given period for imputed lightening data.

These techniques are valuable measurement techniques that are used to compare eight imputation algorithms. RMSE represents the sample standard deviation of the difference between actual and estimated values as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i^{obs} - X_i^{imputed}\right)^2}{n}} \qquad (1)$$

MAE measures the average magnitude of the errors in a set of prediction as:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|X_i^{obs} - X_i^{imputed}\right| \qquad (2)$$

Where n denotes the number of test samples, X_i^{obs} represents the ith target value, $X_i^{imputed}$stands for the predicted value for the ith test sample.

RMSE and MAE both indicate how close the modelled and observed values are. RMSE takes the square root of the average square error, it gives a relatively high weight to the large errors. Therefore, it is appropriate when penalizing large errors are desirable.

B. Estimation Process

The process of the analysis is depicted in Figure 2. Due to the large size of the original dataset, from the original dataset with one minute intervals, the half-hourly dataset is generated based on the average of each 30 minute data and called calibrated dataset. Considering the assumption of “Missing Completely at Random” (MCAR), the percentage of 10%, 20% and 30% missing data were generated from the calibrated dataset. Afterward, missing data were imputed using the 8 methods. In the next step, the difference between the substituted values and real values was computed by RMSE and MAE methods. To provide more accurate comparison, the missing value generation step and the corresponding imputation algorithms were performed for 1000 simulations and the average of the 1000 simulations were used for the final evaluations.

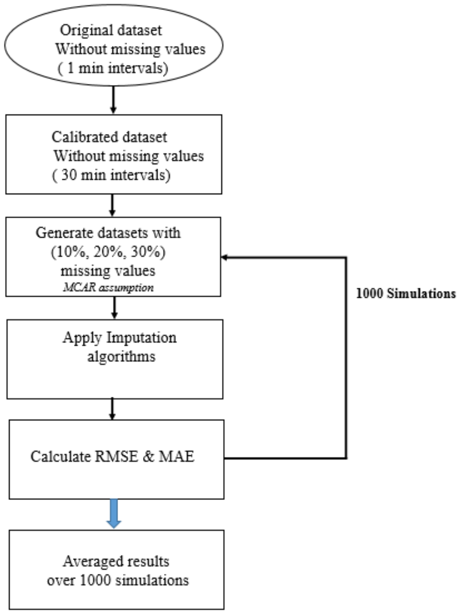
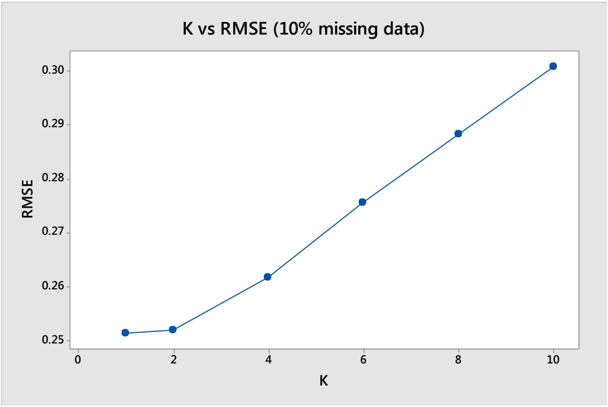


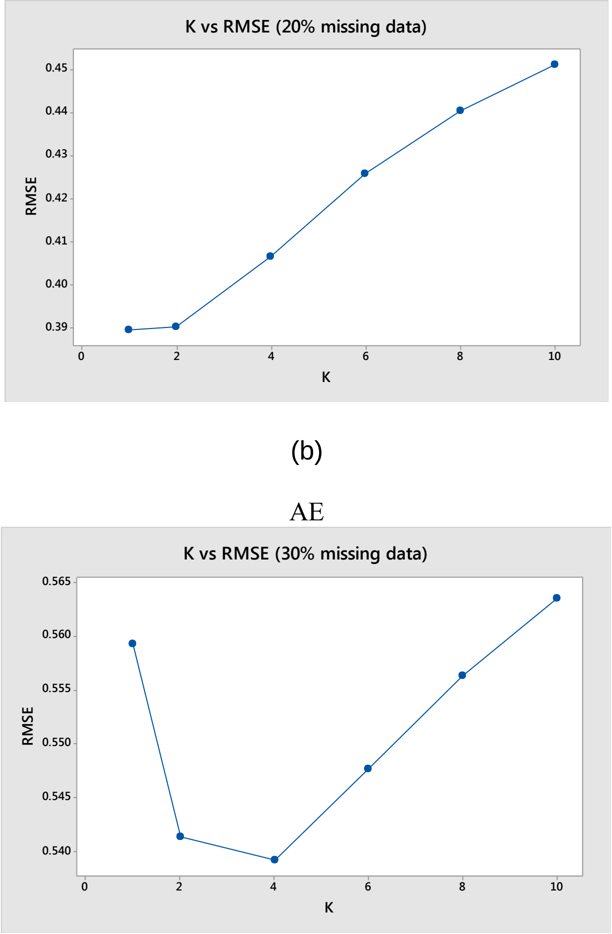
Figure 2Principle of the analysis

C. Result Analysis

As it was mentioned before, in KNN method the number of selected neighbours play an important role. By increasing the percentage of missing data, bigger K value is suitable for the best KNN results. In other word, when the missing data is about 10%, the closest value, to the missed data, is the best value for imputation (Figure 3(a)). However, by increasing the missing data, i.e. for 20% missing, the most optimized K could achieve by considering the K value as 2 or in other word, by considering an hourly boundary, the best value achieved (Figure 3 (b)). For the 30% missing dataset, the best K was 4 which means the boundary of 2 hours could result in better imputation of missing data (Figure 3 (c)). The trend of best K value in terms of missing percentage is depicted in Figure 4. From this figure, it is also obvious that increasing the percentage of missing data results in higher RMSE value which can be considered as a logical confirmation of the principle of our analysis.



(a)



(b)

AE

(c)

Figure 3: relationship between missing percentage and best K value

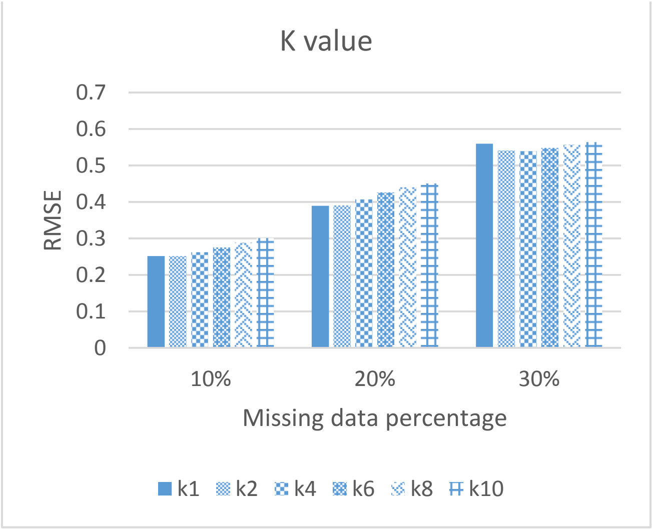


Figure 4: Trend of K values

Figure 5 illustrates the comparison of all methods in terms of computed RMSE.

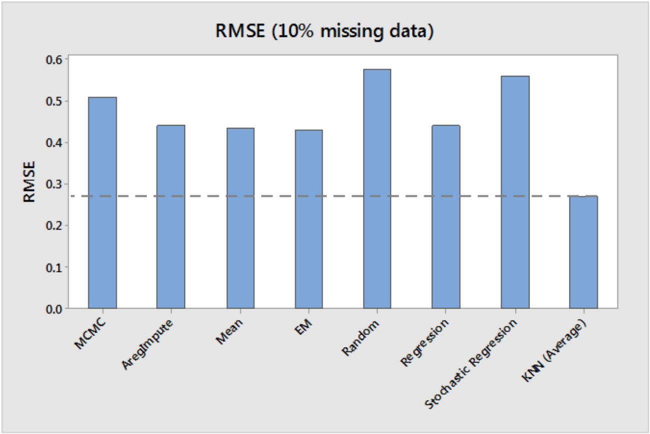
It should be mentioned that to simplify the evaluation, for KNN method, the average of RMSE for each set of missing data (10%, 20% and 30%) is considered for this comparison.

For 10 percent missing data (Figure 5(a)) , Random and Stochastic regression and MCMC techniques achieved the highest percentage of error based on root mean square analysis. With a remarkable gap, KNN shows less error than other methods. AregImpute, Mean, regression and EM techniques achieve the same RMSE , approximately.

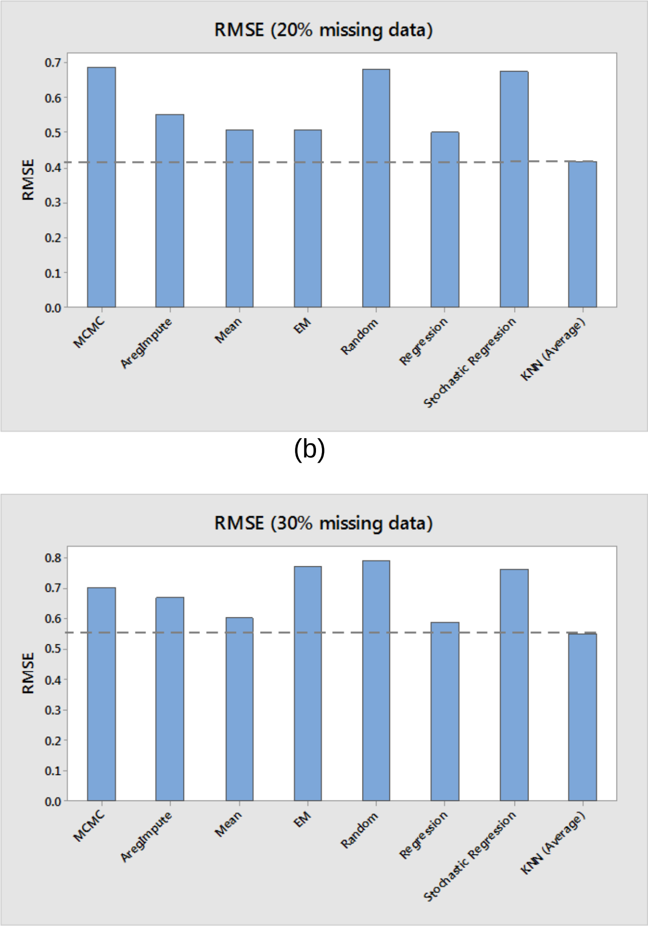
For 20 percent missing values in the dataset (Figure 5(b)), approximately, the same manner in terms of the RMSE values achieved. KNN shows the best and MCMC, Random and Stochastic regression methods achieve the worst methods. RMSE value for AregImpute technique, slightly increased compare with Mean, regression and EM methods.

For 30 percent missing value dataset (Figure 5(c)), KNN, Regression and Mean techniques show the most suitable methods while higher percentage of missing values are available. There is a significant error increase for EM algorithm in this dataset. The KNN and random methods show the best and approximately the worst methods, respectively.

The evaluation of Mean Absolut Errors are depicted in figure 6. Figure 6(a), shows that KNN has a remarkable less error than the other methods. The computed MAE for 20 percent missing data set (Figure 6(b)) and 30 percent missing data (Figure (6(c)) show that the KNN technique archives the lowest error.



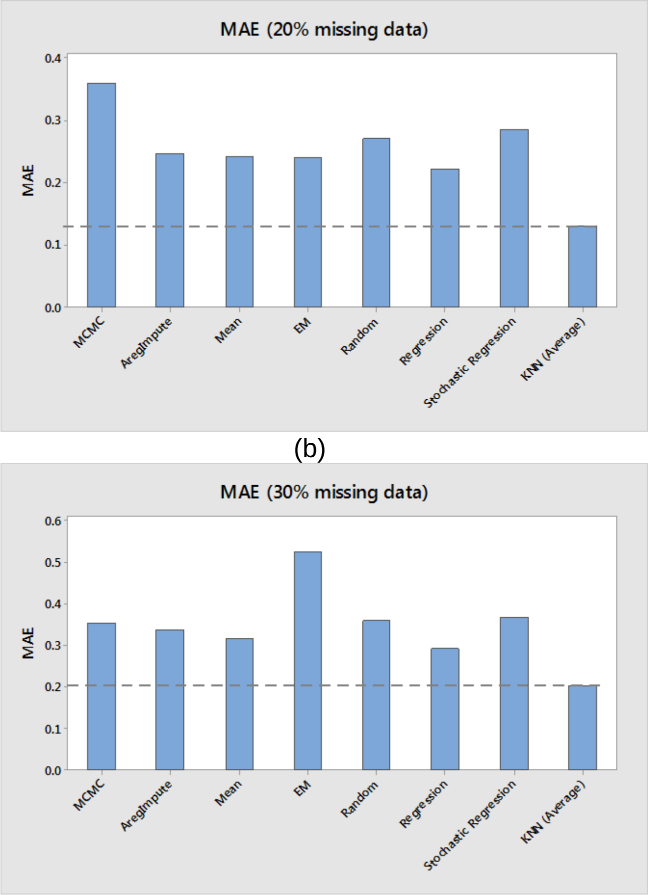
(a)



(b)

(c)

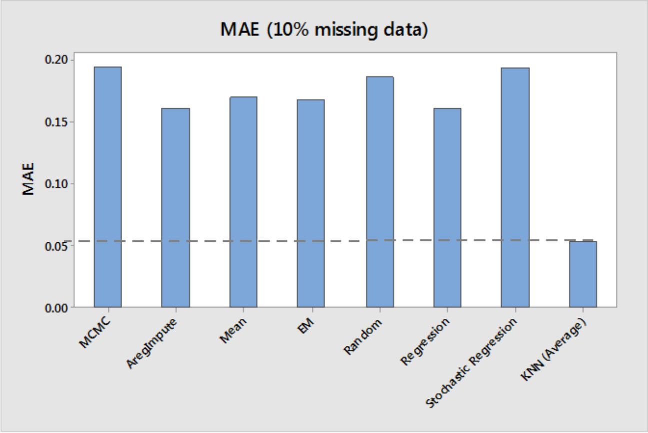
Figure 5 Comparison of methods based on RMSE



(b)

(c)

Figure 6 Comparison of methods based on MAE



(a)

IV. EVALUATION AND OUTCOME

The objective of this research is to highlight the importance of the method that will be used in energy and building fields to treat the missing values. This paper shows that it is important to identify the percentage of missing data before selecting the proper method. In this research eight popular imputation techniques are used on the generated datasets with 10, 20 and 30 percent missing values. The results show that for 10% missing data, KNN achieves a better accuracy in prediction of missing values. Moreover, the best value for K (number of neighbours) find out as one or two which means in this research the best value to be used for replacing missing data for 10 percent data set is the next 30 minutes or next hour of the recorded data.

For the 20% missing data, KNN shows the best results again. In this dataset, it is also concluded that the best value for K is the next 30 minute or next hour to fill the missing data.

For the data set with 30% missing data, KNN again archives the best result. However, the best value for K increased to 4 which means the next two hours of data would be more suitable to be used for the current missing data.

Therefore, it is concluded that increasing the percentage of missing data, requires more neighbours to estimate the missing data.

Additionally , the results of this study showed that the lighting data are more depends on the time instead of the other variables like occupancy. One reason that authors find out is due to the topology of the sensors. The test bed area was equipped with seven occupancy sensors but only one lighting meter. Therefore, the value of occupancy that was used for the imputation, was the average of this data in each 30 minutes interval.

The achievement of this research is limited to the lighting variable, which is strongly time-dependent. In future, we will further investigate other parameters in buildings. Also, the type of the tested building is an educational building. Further investigations are required for other types of building.

ACKNOWLEDGEMENT

The research reported in this paper was supported by Building as a Power Plant: The use of buildings to provide demand response project, funded by the Engineering and Physical Sciences Research Council under Programme Grant EP/P034241/1, and the Active Building Centre (ABC), supported by Industrial Strategy Challenge Fund under Programme Grant EP/S016627/1.

V. References

[1] US. Department of Energy, “EnergyPlus:Engineering Reference,” 2016.

[2] A. Chong and K. Lam, “Uncertainty analysis and parameter estimation of HVAC systems in building energy models,” in *14th Conference of International Building Performance Simulation Association*, Hyderabad, India, 2015.

[3] Y. Heo, R. Choudhary and G. Augenbroe, “Calibration of building energy models for retrofit analysis under uncertainty,” *Energy and Buildings,* vol. 47, pp. 550-560, 2012.

[4] P. Raftery, M. Keane and J. O’Donnell, “Calibrating whole building energy models: An evidence-based methodology,” *Energy and Buildings,* vol. 43, no. 9, pp. 2356-2364, 2011.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman, “Missing value estimation methods for DNA microarrays,” *Bioinformatics,* vol. 17, no. 6, pp. 520-525, 2001.

[6] H. Lewis, “Missing Data in Clinical Trials,” *The New England Journal of Medicine,* vol. 367, pp. 2557-2558, 2012.

[7] T. Schneider, “Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values,” *Journal of climate,* vol. 14, no. 5, pp. 853-871, 2001.

[8] A. Gelman and J. Hill, Data analysis using regression and multilevel/hierarchical models, Cambridge university press, 2006.

[9] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. Brown and R. Pendyala, “Machine learning approaches for estimating commercial building energy consumption,” *Applied energy,* vol. 208, pp. 889-904, 2017.

[10] D. Cabrera and H. Zareipour, “Data association mining for identifying lighting energy waste patterns in educational institutes,” *Energy and Buildings,* vol. 62, pp. 210-216, 2013.

[11] F. Nelwamondo, S. Mohamed and T. Marwala, “Missing data: A comparison of neural network and expectation maximization techniques,” *Current Science,* vol. 93, pp. 1514-1521, 2007.

[12] F. Harrell, “ Hmisc (v 3.0-12): Harrell miscellaneous library for R statistical software,” R package (v 2.2-3), 2006.

[13] J. Leek, E. Monsen, A. Dabney and J. Storey, “ EDGE: extraction and analysis of differential gene expression,” *Bioinformatics,* vol. 22, no. 4, pp. 507-508, 2005.

[14] C. Musil, C. Warner, P. Yobas and S. Jones, “A comparison of imputation techniques for handling missing data,” *Western Journal of Nursing Research,* vol. 24, no. 7, pp. 815-829, 2002.

[15] D. Schunk, “A Markov chain Monte Carlo algorithm for multiple imputation in larg surveys,” *Advances in Statistical Analysis,* vol. 92, pp. 101-114, 2008.

[16] T. Cover and H. Peter, “Nearest neighbor pattern classification,” *IEEE transactions on information theory,* vol. 13, no. 1, pp. 21-27, 1967.

[17] J. Schafer, Analysis of incomplete multivariate data, New York: Chapman and Hall/CRC, 1997.