



Systematic Statistical Analysis to Ascertain the Missing Data Patterns in Energy Consumption Data of Smart Homes

K. Purna Prakash*, Y. V. Pavan Kumar#

*School of Computer Science and Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

#School of Electronics Engineering, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

(kasaraneni.19phd7020@vitap.ac.in, pavankumar.yv@vitap.ac.in)

✉Corresponding Author; Y. V. Pavan Kumar, VIT-AP University, Amaravati-522237, Andhra Pradesh, INDIA

Tel: +91863-2370155, pavankumar.yv@vitap.ac.in

Received: 29.04.2022 Accepted: 23.06.2022

Abstract - The evolution of smart homes is very rapid and the benefits, comfort, as well as flexibility in controlling energy consumption, attract the development of smart home culture across the globe. The energy consumption data collected from these smart homes play a major role in energy pricing, understanding consumers' behaviour, demand-side management, etc., functionalities. But, sometimes, this collected data may suffer from the anomalies such as missing data, redundancy, outliers, etc., which affect the energy data analytics. Among these anomalies, the missing data is one of the anomalies to be concentrated more as it makes data incomplete and significantly hinders the further analysis of the data. This missing of data may take place in three different patterns viz. missing completely at random, missing at random, and missing not at random. Therefore, capturing the pattern of the missing data is highly preferred to better handle them. Although there are a few works on the missing data, they are focused only on the occurrence behaviour, impacts, recovery, and imputation of the missing data rather than identifying the pattern of missing data. Hence to address this problem, this paper proposes a statistical approach to ascertain the pattern of missing data in the energy consumption data of smart homes. The proposed statistical approach revealed that the data are missing at random in the energy consumption data. An energy consumption database named 'Tracebase' is used for implementing the proposed approach.

Keywords Energy consumption data; MAR; MCAR; Missing data pattern; MNAR; Smart home data; Statistical analysis.

1. Introduction

Smart grids/homes and their technologies fulfil the needs of inhabitants by monitoring and controlling in-house actions with self-governance capability, thereby offer them a high standard of living [1-3]. These smart homes are equipped with advanced metering infrastructure to automatically capture the bidirectional data that flow in the power network [4-6]. This household metering infrastructure collects energy consumption data in huge amounts from various smart appliances connected to this power network in smart homes [7, 8]. It is highly desired that this collected data should be error-free to carry out more fruitful analytics [9]. Further, this data should be secured to preserve the privacy of consumers [10]. Thus, the privacy and security of the data

are the elementary concepts of smart homes [11]. Sometimes, this data exhibit abnormal behaviour with missing values, which makes the data incomplete and further affect the analytics. This abnormality in the data is referred to as an anomaly. The analysis of anomalies in energy consumption data has gained prime importance in the last few years [12].

The proper identification of anomalies in the smart home data is essential for conducting effective analytics to load forecasting, demand-side management, elude power wastages and further promise the safety of companies [13]-[16]. The challenges involved in the detection of anomalies need to be analyzed to achieve higher detection rates [17]. Besides, detecting the causes of anomalies at an early stage will play a key role in averting the risk of anomalies [18].

Table 1. Literature works on missing data anomalies

Reference	Year	Key Concept	Description of the Works Carried Out
[24]	2022	Missing data behaviour study	Discussed the occurrence nature of missing data during a day in energy consumption data of smart homes.
[25]	2021	Descriptive analysis of missing data	Presented a descriptive analysis and graphical representation of missing data in the energy consumption database of smart homes.
[26]	2022	Impact of missing data	Discussed the effect of missing data on the performance degradation rate in the photovoltaic systems.
[27]	2020		The assessment of spatial-temporal transient stability in the power systems was presented when there were missing data.
[28]	2016		The profile categorization of AMI load during the existence of missing data.
[29]	2021	Reconstruction of missing data	A least-squares approximated networks method was presented for reconstructing the missing data in power systems' measurement data.
[30]	2019	Recovery of missing data	A regression approach was implemented to recover the missing data in the IoT smart system.
[31]	2019		A novel and robust information-theoretic framework was implemented for assessing the performance of missing data restoration techniques in distributed electrical systems.
[32]	2021	Imputation of missing data	A novel algorithm was implemented to impute the missing data in the smart grids.
[33]	2021		A mixture factor analysis method was implemented to estimate and impute the missing data in buildings' energy consumption data.
[34]	2020		A de-noising auto encoder-based framework was proposed to impute the missing values in the meter data.
[35]	2020		A bi-directional imputation scheme for missing data relied on a long/short-term memory technique that was implemented on the building's energy consumption data.
[36]	2020		A novel clustering strategy viz., correlation clustering imputation was discussed to estimate the missing data and its imputation for power grid application.
[37]	2020	Missing data pattern	The classifications of missing data viz., missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), and the application of linear regression to find missing data pattern was discussed. Also, studied the effect of missing data on wind farm time-series data forecasting.
[38]	2017	Missing data handling	A fuzzy inductive reasoning technique was presented to deal with missing data of smart grids for the purpose of forecasting.

A framework was implemented using huge volumes of smart meter data to detect anomalies in smart grids [19]. The data quality issues and their classification viz. missing data, noisy data, and outliers were discussed in [20]. Further, a framework was discussed in [21] to detect anomalies and faulty meters in smart grids. The detection of the topology errors in the power grids using hypothesis testing was discussed in [22]. The possible opportunities and issues related to big data in electric utilities were discussed in [23]. In addition to the above, key state-of-the-art works on missing data analysis, which leads to the considered problem statement in this paper are discussed in Table 1.

The abovementioned works focused on the occurrence behaviour, impacts, recovery, and imputation of the missing data rather than the detection of missing data patterns in the energy consumption data of smart grids/buildings/homes. Although there was work on finding missing data patterns using linear regression [37], it is not sufficient alone to find the missing data patterns. Hence, this paper proposes a systematic statistical approach to ascertain the missing data pattern in the energy consumption database of smart homes, which is the main contribution of this paper. By considering the gaining importance of statistical methods in analyzing the energy consumption data [39], the proposed study in this paper implements various key statistical methods, as discussed in Section 2.

Other sections of the paper are arranged as follows. In Section 2, the description and implementation of various key statistical methods for missing data pattern identification are explained. In Section 3, the results are projected and a clear analysis of the observations is provided; and the conclusions are highlighted in Section 4.

2. Description and Implementation of the Proposed Approach

This section discusses the conceptual model and the implementation of the suggested statistical approach. The conceptual model for the proposed approach is presented in Fig. 1, where, the energy consumption database of smart homes is given as the input. The proposed approach consists of two major steps viz., fill missing timestamps and place "not available (NA)" values, and identify missing data patterns using statistical methods. To identify the missing data pattern, several statistical methods such as Little's MCAR test, logistic regression, normality testing, and t-test/Wilcoxon rank-sum test are used. The implementation of these statistical methods results in the missing data pattern that exists in the energy consumption database.

The implementation flow of the suggested statistical approach is presented in Fig. 2. This approach is implemented by considering the energy consumption data of a device with the identifier "dev_768D06" connected on "20/05/2012" in the appliance "MicrowaveOven". Out of 43 appliances in the "Tracebase" dataset [40], this appliance exhibited the highest number (84740) of missing data instants on 20/05/2012 [24]. Hence, this data is considered as input to implement the proposed approach. The proposed approach is implemented in two parts. The first part deals with the data preparation, filling of missing timestamps, and placing NA values in the respective timestamps as discussed in Subsection 2.1. The second part deals with ascertaining missing data patterns as discussed in Subsection 2.2.

Further, the description of various statistical tests used to ascertain the missing data pattern is given in Subsection 2.3. This paper is mainly focused on the details and description of the statistical tests than the mathematical foundations. So, the

concepts of Little’s MCAR test, logistic regression, normality testing, t-test, and Wilcoxon rank-sum test are

discussed in subsections 2.3.1 through 2.3.5. This proposed statistical analysis is implemented using RStudio IDE.

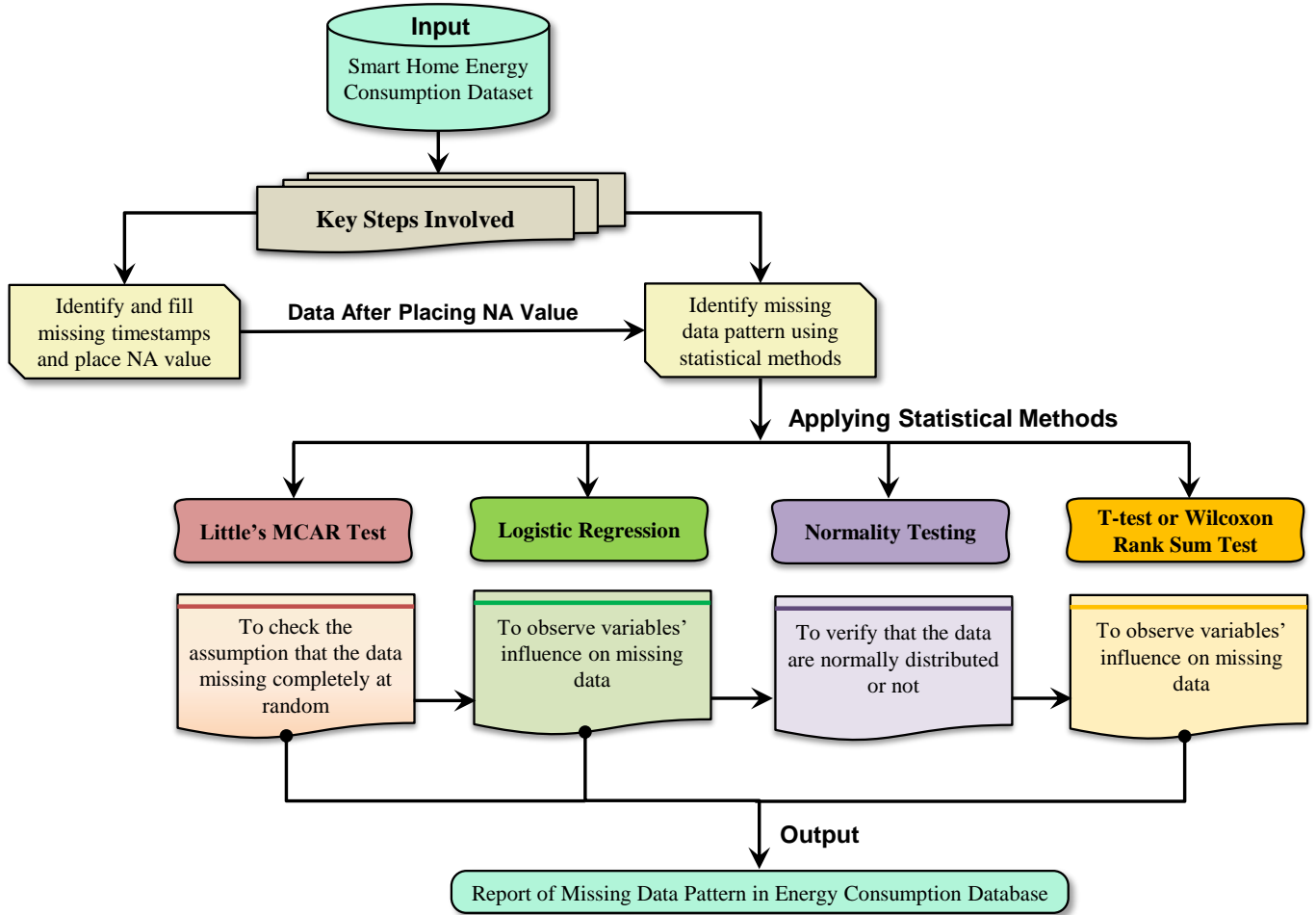


Fig. 1. Conceptual model of the proposed approach.

2.1. Filling Missing Timestamps and Placing NA

The process in the proposed statistical analysis starts with the loading of directories and comma-separated values (CSV) files from the considered data input. The information of these directories and files is saved in the objects “dir_info” and “file_info” respectively. Each CSV file is read from these objects by using “for(p in 1:length(dir_info))” and “for(q in 1:length(file_info))”. The current state of the data is not suitable for applying the proposed analysis directly. Hence, the data in the CSV file is split into the required columns, namely, “TRACED_DATETIME” that consists of timestamp information and “TRACED_READING” that consists of the energy consumption reading information.

The “TRACED_DATETIME” column is converted into the “POSIXlt” date format for making the timestamp information flexible for filling missing timestamps. Further, the date information from the “TRACED_DATETIME” is extracted and converted into “date” format. This information is saved in the object “date_info”. To complete this filling process, the start time and end time of the day should be considered. Hence, the start time sequence and end time sequence is declared as “start_time_seq = 00:00:00” and “end_time_seq = 23:59:59” for “date_info”. The time sequence in the “TRACED_DATETIME” is verified based

on the seconds information by using the keyword “sec” in the function “complete(TRACED_DATETIME = seq(start_time_seq, end_time_seq, by = ‘sec’))”. If the “TRACED_DATETIME” is NULL, the missing timestamp is filled and NA is placed in the respective position of the “TRACED_READING” column. If it is not NULL, then the next timestamp will be verified. For further study, the column “TRACED_DATETIME” is split into the columns namely, “TRACED_DATE”, “TRACED_HOUR”, “TRACED_MINUTE”, and the “TRACED_SECOND”. All these columns’ information is saved into a new CSV file that is used for the analysis.

2.2. Ascertaining Missing Data Pattern

The new CSV file is read and explored to visualize the missing data in each column. The process is continued with some statistical tests to ascertain the missing data pattern in the energy consumption data. Initially, a statistical test named “Little’s MCAR” test is applied to verify the data are MCAR. The outcome (p-value) of this test reveals whether the data are MCAR or not. If the p-value is lesser than 0.05 (statistically noticeable) then the data pattern is not MCAR, which denotes that there is a relation between the columns, while the data are missed. If the p-value is more than 0.05

then the data pattern is MCAR, which denotes that there is no relation between the columns while the data are missed.

If the data are not MCAR, the other statistical tests such as logistic regression and t-test are applied to observe whether the data are missing at random. To conduct these tests, a new column "STATUS" needs to be added to the new CSV file, which includes the encoded information of missing data status (i.e., missing = 1, not missing = 0) in the "TRACED_READING" column. This "STATUS" column is further used in the prescribed statistical tests. Once the data are ready, the logistic regression is applied to the other columns using the "STATUS" column. The outcome (p-value) of the logistic regression reveals whether the data are MAR or not. The columns with a p-value lesser than 0.05 are

statistically noticeable and they have a significant influence on the missing data, which represents that the data are MAR.

Before applying the t-test, it is recommended to verify whether the data are normally distributed or not by conducting the normality test. If the data are normal, a t-test is conducted. If the data are non-normal, then a non-parametric test, named "Wilcoxon rank-sum test" is conducted. This test is equivalent to the t-test and is applied to all the other columns by using the "STATUS" column. From the outcomes of these tests, the columns that consist p-value less than 0.05 are statistically significant and influence the missing data. Subsequently, the data are said to be MAR. The process is stopped once the missing data pattern is ascertained.

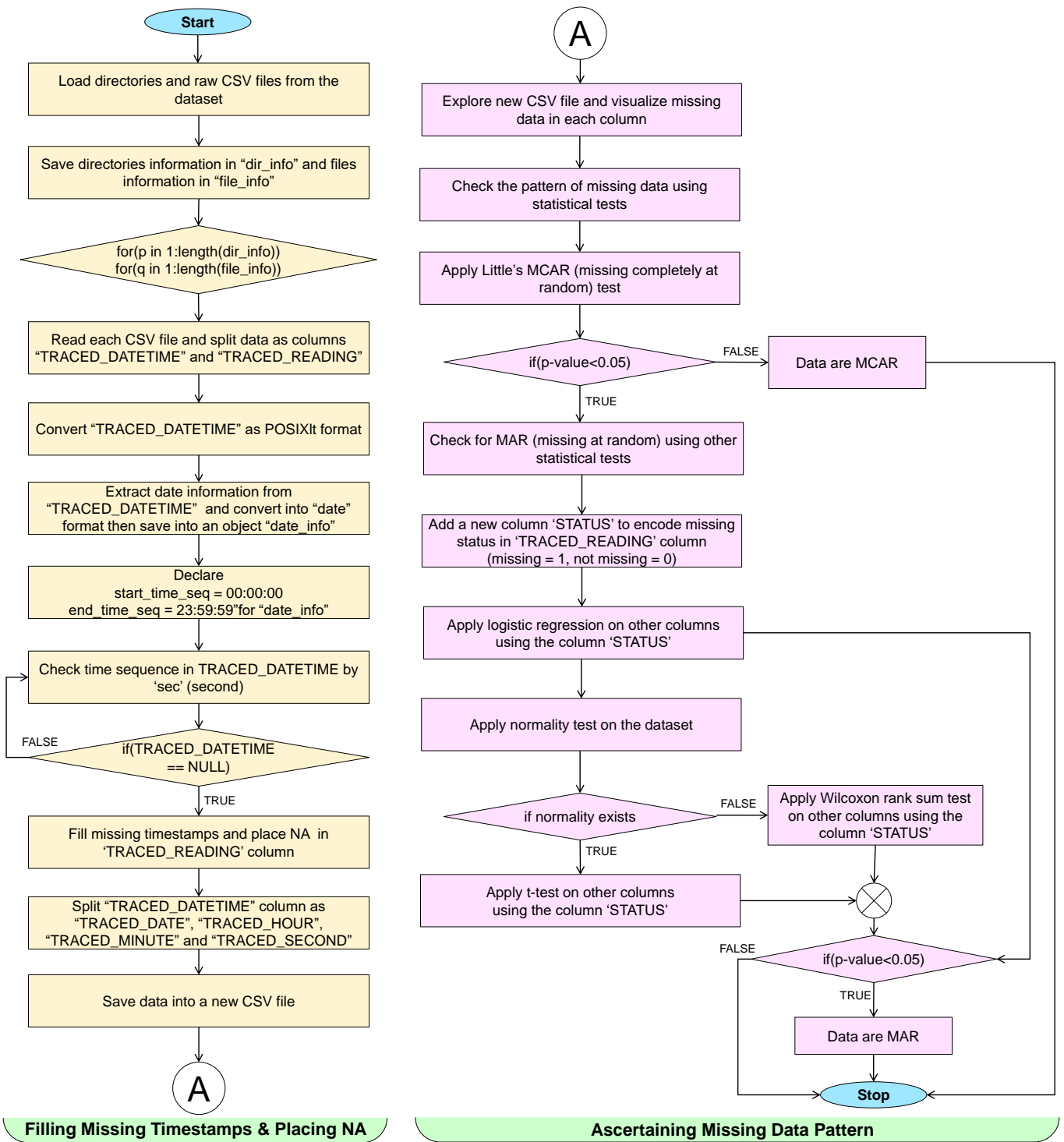


Fig. 2. Flowchart for the proposed statistical approach.

2.3. Description of Statistical Tests

2.3.1 Little's MCAR Test

This test was developed by Little in 1988 to test the assumption that the data are MCAR [41]. This is a hypothesis test, where the null hypothesis is that the data are MCAR. To implement this test, a function "mcar_test()" that works on the library "naniar" can be used in RStudio IDE. The outcome of this test will contain the values such as statistic (a chi-squared value), degrees of freedom (df) for the

chi-squared statistic value, probability value (p-value), and missing patterns, which represent the number of missing data patterns existing in the data. If the output of this test results in a high statistic value and low p-value (i.e., lesser than the critical value, 0.05), then the null hypothesis will be ignored and the data are not MCAR. This represents that there is a correlation between the variables while the data are missing [42].

2.3.2 Logistic Regression

This is a better approach to deal with the binary dependent variables, where the case that, there are two categories such as yes or no, 1 or 0, etc. [43]. The logistic function was developed by French mathematician Pierre François Verhulst in the 19th century. The main theme of this function is to test whether an event occurred or not irrespective of that when it occurred. To implement this test, the dependent variable should be categorical, and the independent variables need not be normal, linear, interval, and of equal variance. The values in the dependent variable will be encoded with 1 and 0, when there is a response encoded as 1 and when there is no response encoded as 0 [44].

The functionality of the logistic regression is similar to linear regression, but with a binomial response [45]. The variables that are statistically significant (i.e., p-value lesser than the critical value, 0.05) will influence the missing data. This test can be implemented by using a generalized linear model i.e., “glm()” function with family type “binomial” in RStudio IDE.

2.3.3 Normality Testing

There are numerical and graphical methods to test whether the given sample is distributed normally or not [46]. The quantile-quantile plot is one of the visual methods and an alternative to the histograms to test the normal distribution. In short, the quantile-quantile plot is also referred to as the QQ plot. A straight line will be generated and it goes through the origin if the distributions are the same, otherwise, the position of the straight line will be changed. If the data points lie closer to the straight line, then the data are normally distributed. If the data points are far from the straight line, then the data are non-normal [47].

The QQ plot is a more useful and accurate visual method to check the normal distribution very easily when dealing with large sample sizes [48]. This test can be implemented by using qqPlot() function in RStudio IDE.

2.3.4 Independent T-Test

A t-test is a parametric test, which is also denoted as a student's t-test. It was designed in 1908 by William Sealy Gosset. It is applied to test the key difference between the mean values of two independent groups. This test will be applied mostly when the data are normally distributed and not suitable for the samples of small size and when the data are not normally distributed [49]. This test can be implemented using the t.test() function in RStudio IDE.

2.3.5 Wilcoxon Rank-sum Test

This is a non-parametric test, which is a substitute to the t-test that can be applied when data are non-normal. The scientists, Henry Mann and Donald Ransom performed a detailed analysis of the statistic in the year 1947. Hence, this test is also denoted as the “Mann-Whitney U test and Wilcoxon-Mann-Whitney” test [50]. This test can be validated by using the obtained p-value. The variables that are statistically significant (i.e., p-value lesser than the critical value, 0.05) will influence the missing data. When the statistical tests provide the information of significance, then it represents that there is a correlation between the variables and impacts the missing data. This infers that the data are MAR and not MCAR [51].

3. Results and Analysis

The results of the suggested statistical approach are presented in three subsections. In subsection 3.1, the results related to filling missing timestamps and placing NA values are discussed. In subsection 3.2, the visualizations of missing data are discussed. In subsection 3.3, the results related to checking the type of missing data pattern are discussed.

3.1. Results of Filling Missing Timestamps and Placing NA Values

The data in the raw CSV file of the device “dev_768D06” connected on 20/05/2012 in the appliance “MicrowaveOven” is shown in Fig. 3(a). From this, it is observed that the data are stored in a single column format and the type of this column's data is a string. This data provides information about energy consumption data, which is a mixture of the “timestamp” and “reading” information. However, understanding of these information directly from Fig. 3(a) is very difficult. So, it is rearranged shown in Fig. 3(b). Here, the predefined single column data is split into various columns, viz., TRACED_DATE, TRACED_HOUR, TRACED_MINUTE, TRACED_SECOND, and TRACED_READING.

Further, it can be observed that the data available from hour 14 and the data for previous hours are missed. But, the actual scenario for data capturing is that it will be done for all 24 hours where it starts from 00:00:00 and ends at 23:59:59. Although, the data for hour 14 are available, still the missing data is observed in several seconds. Hence, those missing timestamps are filled and the respective reading information placed in the TRACED_READING column with NA value as shown in Fig. 3(b).

	A	B	C
1	20/05/2012 14:46:07;0;0		
2	20/05/2012 14:46:09;0;0		
3	20/05/2012 14:46:19;0;0		
4	20/05/2012 14:46:28;0;0		
5	20/05/2012 14:46:29;0;0		
6	20/05/2012 14:46:31;0;0		
7	20/05/2012 14:46:34;0;0		
8	20/05/2012 14:46:48;0;0		
9	20/05/2012 14:47:02;2;0		
10	20/05/2012 14:47:11;0;0		
11	20/05/2012 14:47:20;0;0		
12	20/05/2012 14:47:31;0;0		
13	20/05/2012 14:47:58;0;0		
14	20/05/2012 14:48:10;0;0		
15	20/05/2012 14:48:26;0;0		
16	20/05/2012 14:48:36;0;0		
17	20/05/2012 14:48:49;0;0		
18	20/05/2012 14:48:50;0;0		
19	20/05/2012 14:49:02;0;0		
20	20/05/2012 14:49:15;0;0		
21	20/05/2012 14:49:27;0;0		
22	20/05/2012 14:49:39;0;0		
23	20/05/2012 14:49:51;0;0		
24	20/05/2012 14:50:30;2;0		
25	20/05/2012 14:50:31;2;0		

a) Raw CSV file

	A	B	C	D	E
1	TRACED_DATE	TRACED_HOUR	TRACED_MINUTE	TRACED_SECOND	TRACED_READING
2	20-05-12	0	0		0 NA
3	20-05-12	0	0		1 NA
4	20-05-12	0	0		2 NA
5	20-05-12	0	0		3 NA
6	20-05-12	0	0		4 NA
7	20-05-12	0	0		5 NA
8	20-05-12	0	0		6 NA
9	20-05-12	0	0		7 NA
10	20-05-12	0	0		8 NA
11	20-05-12	0	0		9 NA
12	20-05-12	0	0		10 NA
13	20-05-12	0	0		11 NA
14	20-05-12	0	0		12 NA
15	20-05-12	0	0		13 NA
16	20-05-12	0	0		14 NA
17	20-05-12	0	0		15 NA
18	20-05-12	0	0		16 NA
19	20-05-12	0	0		17 NA
20	20-05-12	0	0		18 NA
21	20-05-12	0	0		19 NA
22	20-05-12	0	0		20 NA
23	20-05-12	0	0		21 NA
24	20-05-12	0	0		22 NA
25	20-05-12	0	0		23 NA

b) New CSV file

Fig. 3. CSV file of energy consumption data on 20/05/2012.

3.2. Results of Visualizing Missing Data in the Dataset

The details of the CSV file data are shown in Fig. 4. From this, it is observed as the proportion of discrete columns is 20% and continuous columns is 80%. And, the other metrics that are calculated are, all missing columns (as 0%), complete rows (as 1.93%), and missing observations (as 19.61%).

The proportions of the missing data in each column after filling the missing timestamps and placing NA values are shown in Fig. 5. From this, it can be observed that the columns TRACED_DATE, TRACED_HOUR, TRACED_MINUTE, and TRACED_SECOND are 0% of missing data. But, the TRACED_READING column has 98.07% of the missing data.

The visualization of the missing data in the considered dataset for all minutes of each hour is referred to in Fig. 6. In this figure, it is observed that the minutes' information of respective hours is taken on the x-axis, and readings information is taken on the y-axis. The red-coloured dots indicate the data that are missing and the blue-coloured dots specify the data that are not missing. Further, it is observed that there is complete missing data in all minutes of the hours from 0 through 13 and the actual recording of data can be

observed from minute 46 of hour 14. To better understand the existence of the missing data, the zoomed version of hour 20 is shown as it seems to have more non-missing data points when compared with other hours in the visualization. In this hour, the range of missing data can be observed in almost all the minutes.

The visualization of the missing data in all seconds of each minute in hour 20 is shown in Fig. 7. From this, it is observed that the seconds' information of respective minutes is taken on the x-axis, and readings information is taken on the y-axis. Further, it is observed that the data are missing in the majority of seconds in the respective minutes of hour 20. To better understand the missing data, the zoomed version of minute 51 of hour 20 is shown. From this, it is observed that the data are missing in the majority of the seconds.

3.3. Results of Ascertaining Missing Data Pattern

The ascertaining of missing data patterns is carried through realization of various statistical tests as mentioned in above sections. The results corresponding to Little's MCAR test, logistic regression, normality testing using Quantile-Quantile (QQ) plots, and Wilcoxon rank-sum test are described through subsections 3.3.1 to 3.3.4 respectively.

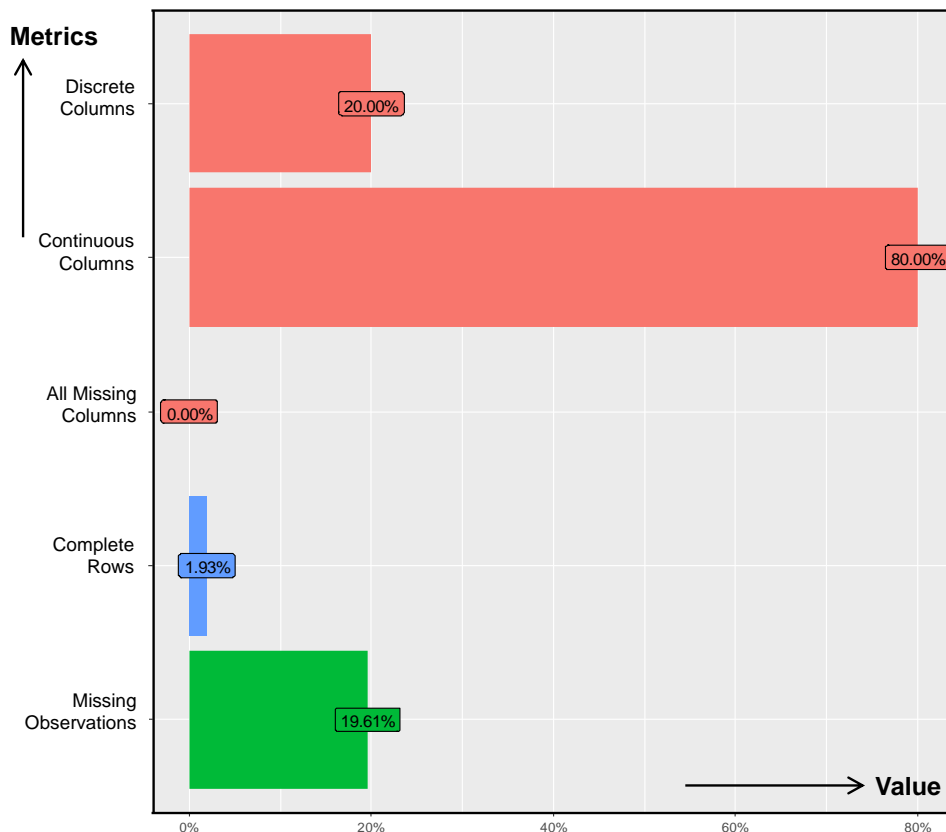


Fig. 4. Details of the CSV file data.

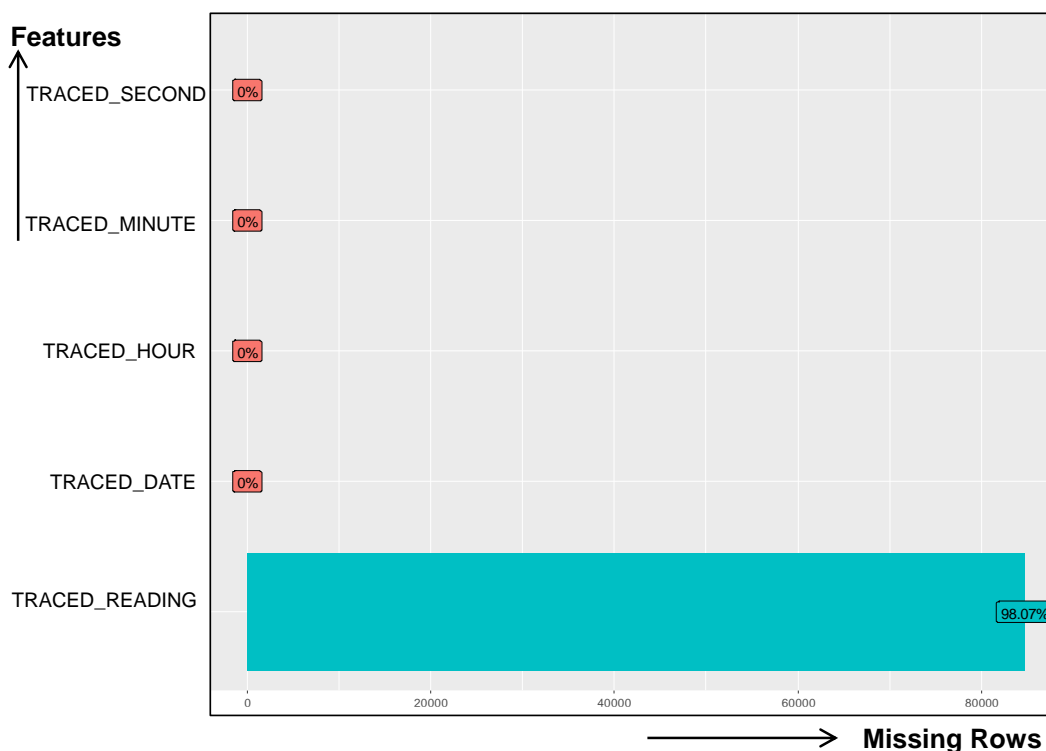


Fig. 5. Proportions of missing data in each column.

3.3.1 Result of Little's MCAR Test

The output of Little's MCAR test is shown in Table 2. This table consists of four columns viz., Statistic, Degree of Freedom, P-value, Missing-patterns. From this, it is observed

that there is a high statistic value (1683) and very low p-value (0), which represents that the data are statistically significant and not MCAR. Thus, this test ascertained that the data are not MCAR. Further, the column "Missing-patterns" showcases that there are two patterns

related to missing data. As the data are not MCAR, the other statistical tests mentioned above need to be performed to verify the existence of patterns MAR or MNAR.

Table 2. Output of Little’s MCAR test

Statistic	Degree of Freedom	P-value	Missing.patterns
1683	2	0	2

3.3.2 Result of Logistic Regression Test

The logistic regression is applied to the columns TRACED_HOUR, TRACED_MINUTE, and

TRACED_SECOND using the column STATUS. The output of this test revealed that the p-values ($<2e-16$, 0.0122) for the columns TRACED_HOUR and TRACED_MINUTE respectively are less than 0.05 and are said to be statistically significant. This represents that these two columns have a significant influence on the missing data. The p-value (0.3719) for the column TRACED_SECOND is greater than 0.05 and statistically not significant. This represents that TRACED_SECOND does not have any influence on the missing data. The information of obtained p-values in the logistic regression is shown in Fig. 8.

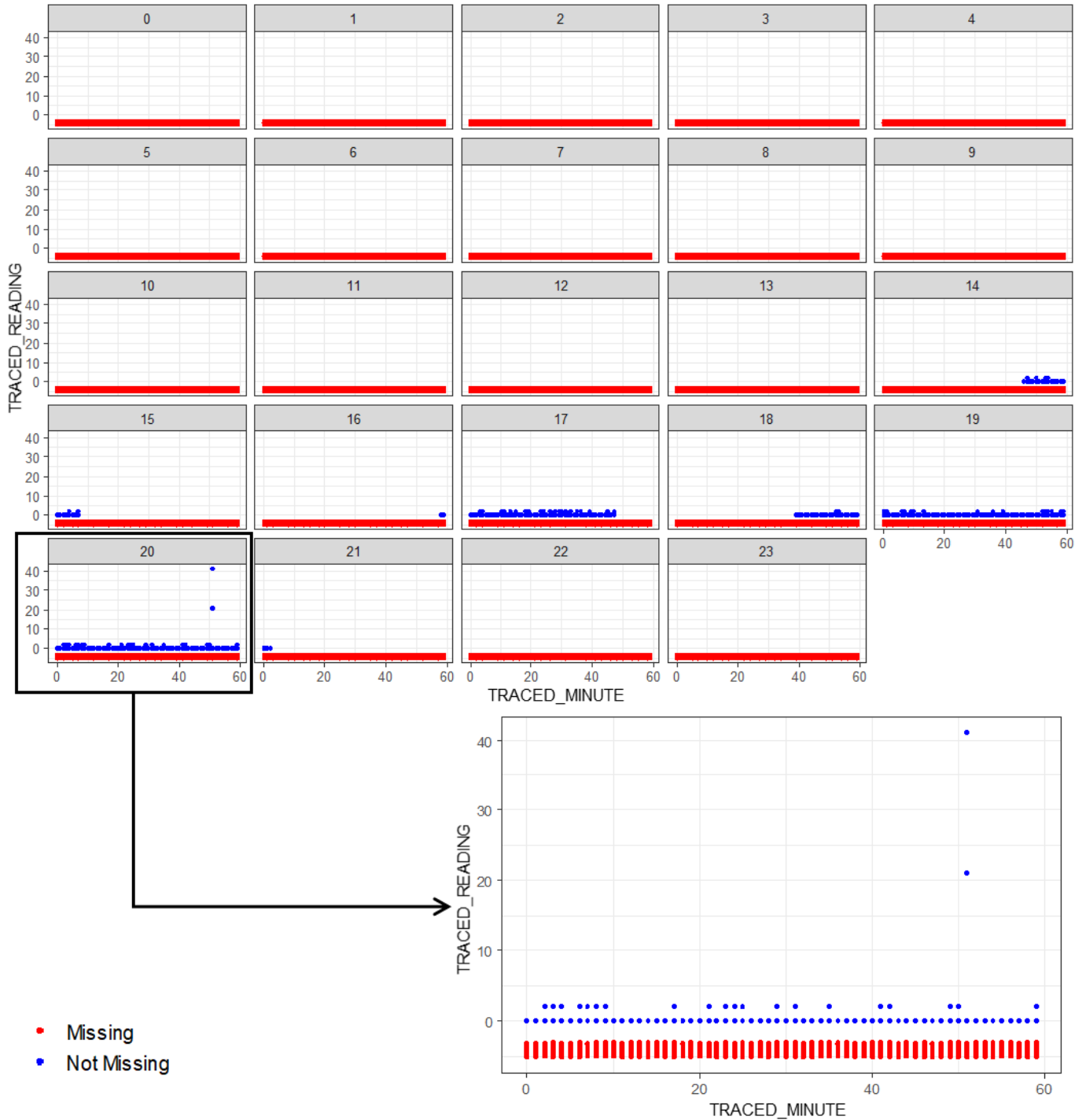


Fig. 6. Visualization of missing data in all minutes of each hour.

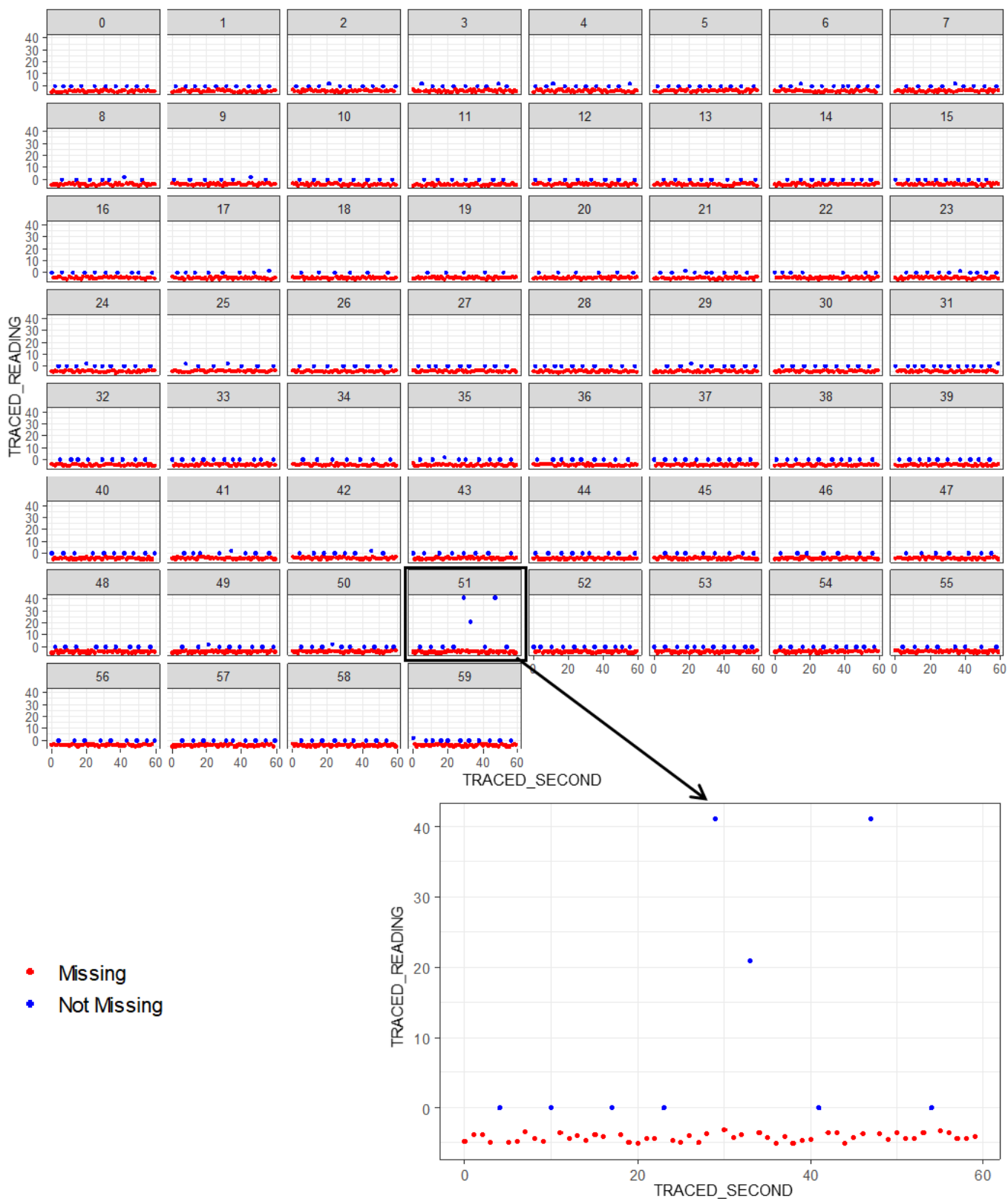


Fig. 7. Visualization of missing data in all seconds of each minute in hour 20.

Output of Logistic Regression

Call:
 glm(formula = STATUS ~ TRACED_HOUR + TRACED_MINUTE +
 TRACED_SECOND,
 family = binomial, data = dataset)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.88710	0.07459	0.13177	0.23541	0.43221

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	6.986963	0.117715	59.355	<2e-16 ***
TRACED_HOUR	-0.190137	0.005306	-35.836	<2e-16 ***
TRACED_MINUTE	-0.003626	0.001447	-2.506	0.0122 *
TRACED_SECOND	-0.001291	0.001446	-0.893	0.3719

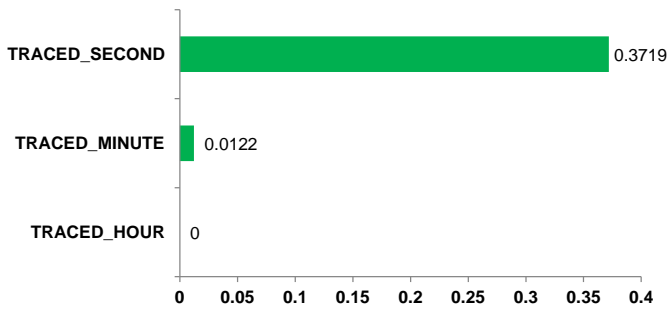


Fig. 8. Obtained p-values in logistic regression.

3.3.3 Result of Normality Test using QQ Plots

The quantile-quantile (QQ) plot is applied to the columns TRACED_HOUR, TRACED_MINUTE, TRACED_SECOND, and TRACED_READING to test whether the data in these columns are normally distributed or not. The visualizations of QQ plots are given in Fig. 9. From this, it is seen that the data in columns TRACED_HOUR, TRACED_MINUTE, TRACED_SECOND, and TRACED_READING are not normally distributed as many

data points are far from the straight line. Hence, the t-test does not apply to non-normal data.

3.3.4 Result of Wilcoxon Rank-sum Test

As discussed in section 3.3.3, the t-test is not applicable because the considered data does not have normality. Thus, the Wilcoxon rank-sum test is applied to the columns TRACED_HOUR, TRACED_MINUTE, and TRACED_SECOND using the column STATUS. The output of this test revealed that the p-values ($< 2.2e-16$, 0.01332) of the columns TRACED_HOUR and TRACED_MINUTE are less than 0.05 and are said to be statistically significant. This represents that these two columns have a significant influence on the missing data.

The p-value (0.3766) for the column TRACED_SECOND is greater than 0.05 and statistically not significant. This represents that the column TRACED_SECOND does not have any influence on the missing data. The information of obtained p-values in the Wilcoxon rank-sum test is shown in Fig. 10.

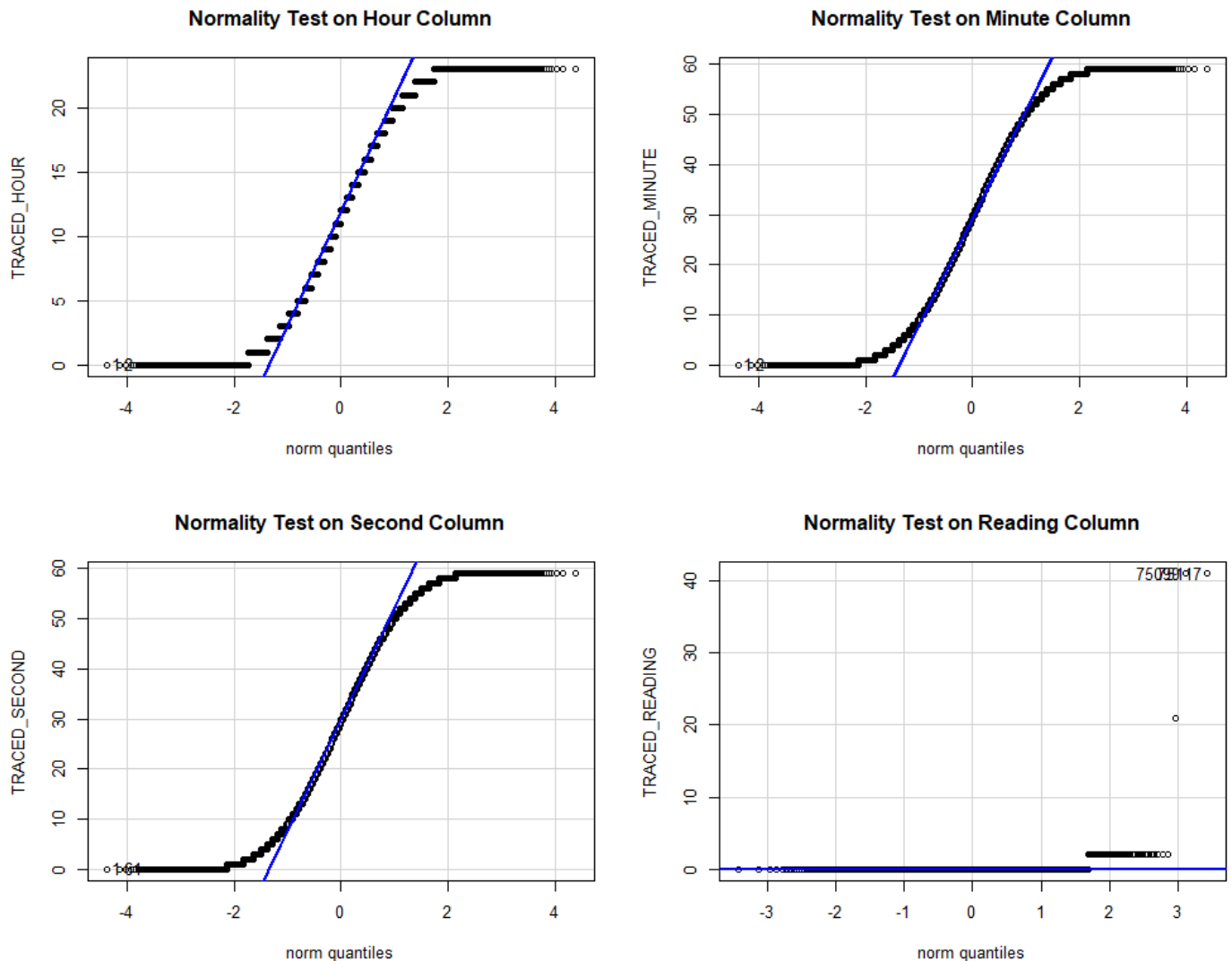


Fig. 9. Quantile-Quantile plots for normality testing.

Output of Wilcoxon Rank-sum Test	
Result-1:	
data: TRACED_HOUR by STATUS	
W = 1.12e+08, p-value < 2.2e-16	
alternative hypothesis: true location shift is not equal to 0	
Result-2:	
data: TRACED_MINUTE by STATUS	
W = 73213085, p-value = 0.01332	
alternative hypothesis: true location shift is not equal to 0	
Result-3:	
data: TRACED_SECOND by STATUS	
W = 71607761, p-value = 0.3766	
alternative hypothesis: true location shift is not equal to 0	

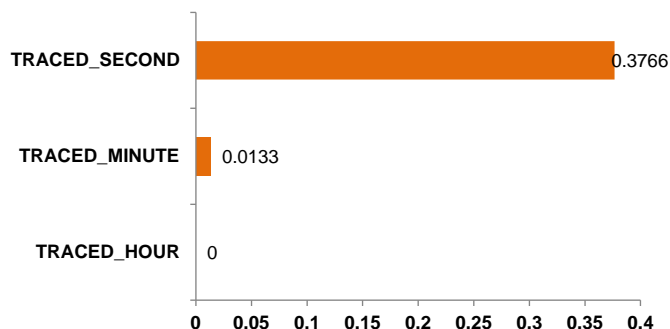


Fig. 10. Obtained p-values in Wilcoxon rank-sum test.

The summary of p-values for logistic regression and the Wilcoxon rank-sum test is presented in Table 3. From this, it can be observed that the columns TRACED_HOUR and TRACED_MINUTE are significant and correlated to the missing data. This correlation indicates that the data in the considered energy consumption database of smart homes are MAR and not MCAR.

Table 3. Summary of p-values in logistic regression and Wilcoxon rank-sum test

Name of the Test	Column Name	P-value	Significant or Not
Logistic Regression	TRACED_HOUR	<2e-16	Significant
	TRACED_MINUTE	0.0122	Significant
	TRACED_SECOND	0.3719	Not Significant
Wilcoxon Rank-sum Test	TRACED_HOUR	< 2.2e-16	Significant
	TRACED_MINUTE	0.01332	Significant
	TRACED_SECOND	0.3766	Not Significant

4. Conclusions

Thus, the proposed statistical analysis successfully ascertained the missing data pattern in the considered energy consumption database of smart homes. The existence of missing data in this database is legibly visualized. Further, all the missing timestamps in the energy consumption data are successfully filled with respective timestamps and the NA value is placed in the READING column respectively. The salient observations made from the implemented statistical tests are given as follows.

- The outcome of Little's MCAR test provided the information that the data are not MCAR.
- The outcome of logistic regression provided information that the columns TRACED_HOUR and TRACED_MINUTE influence the missing data, and

the column TRACED_SECOND do not influence the missing data.

- The outcome of normality testing using QQ plots provided the information that the data distribution is non-normal. Hence, the Wilcoxon rank-sum test is implemented on the data instead of the t-test.
- The outcome of the Wilcoxon rank-sum test provided information that the columns TRACED_HOUR and TRACED_MINUTE influence the missing data and the column TRACED_SECOND do not influence the missing data.

In summary, from the above statistical analysis test results, it is observed that the columns TRACED_HOUR and TRACED_MINUTE have a significant impact on the missing data. Hence, it is concluded that the considered smart home energy consumption data are a type of MAR (missing at random), which are feasible for data imputation.

So, this analysis is recommended to be performed on the energy consumption databases to understand whether that data are feasible for data imputation or not. (i.e., if the considered data is declared as MAR, then it is feasible for imputation).

The presented statistical methods and the analysis in this paper can be extended to all other smart home databases.

Acknowledgements

This work was supported in part by Project Grant No: SRG/2019/000648, sponsored by the Start-up Research Grant (SRG) scheme of the Science and Engineering Research Board (SERB), a statutory body under the Department of Science and Technology (DST), Government of INDIA.

References

- [1]. D. N. Mekuria, S. Paolo, N. Falcionelli, and A. D. Franco, "Smart home reasoning systems: a systematic literature review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4485-4502, 2021.
- [2]. F. Ayadi, I. Colak, I. Garip and H. I. Bulbul, "Targets of countries in renewable energy," 2020 9th International Conference on Renewable Energy Research and Application (ICRERA), Glasgow, pp. 394-398, 27-30 September 2020.
- [3]. I. Colak, R. Bayindir and S. Sagiroglu, "The effects of the smart grid system on the national grids," 2020 8th International Conference on Smart Grid (icSmartGrid), Paris, pp. 122-126, 17-19 June 2020.
- [4]. A. Al-Abri, W. Al Khalil, K. E. Okedu, "Electricity Sector of Oman and Prospects of Advanced Metering Infrastructures," *International Journal of Smart Grid (ijSmartGrid)*, vol. 6, no. 1, pp. 1-12, March 2022.
- [5]. A. Colak, N. Guler and K. Ahmed, "Intelligent communication techniques for smart grid systems: A survey," 2021 9th International Conference on Smart Grid (icSmartGrid), Setubal, pp. 273-277, 29 June 2021 - 01 July 2021.
- [6]. G. Pradeep Reddy and Y. V. Pavan Kumar, "Retrofitted IoT based communication network with hot standby router protocol and advanced features for smart

- buildings,” *International Journal of Renewable Energy Research (IJRER)*, vol. 11, no. 3, pp. 1354-1369, September 2021.
- [7]. Q. Zhao, Z. Chang and G. Min, “Anomaly detection and classification of household electricity data: a time window and multilayer hierarchical network approach,” *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3704-3716, March 2022.
- [8]. S. Sagioglu, R. Bayindir, Y. Canbay, and I. Colak, “Big data issues in smart grid systems,” 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), Birmingham, pp. 1007-1012, 20-23 November 2016.
- [9]. K. P. Prakash and Y. V. P. Kumar, “A systematic approach for exploration, behavior analysis, and visualization of redundant data anomalies in smart home energy consumption dataset,” *International Journal of Renewable Energy Research (IJRER)*, vol. 12, no. 1, pp. 109-123, March 2022.
- [10]. A. Zielonka, W. Marcin, S. Garg, G. Kaddoum, Md. P. Jalil and M. Ghulam, “Smart homes: How much will they support us? A research on recent trends and advances,” *IEEE Access*, vol. 9, pp. 26388-26419, January 2021.
- [11]. J. F. DeFranco and K. Mohamad, “Smart home research themes: An analysis and taxonomy,” *Procedia Computer Science*, vol. 185, pp. 91-100, 2021.
- [12]. W. Zhang, D. Xiaowei, H. Li, X. Jin and D. Wang, “Unsupervised detection of abnormal electricity consumption behavior based on feature engineering,” *IEEE Access*, vol. 8, pp. 55483-55500, 2020.
- [13]. A. Sial, S. Amarjeet and A. Mahanti, “Detecting anomalous energy consumption using contextual analysis of smart meter data,” *Wireless Networks*, vol. 27, pp. 4275-4292, 2021.
- [14]. L. G. Fahad and F. S. Tahir, “Activity recognition and anomaly detection in smart homes,” *Neurocomputing*, vol. 423, pp. 362-372, 2021.
- [15]. G. Fenza, M. Gallo, and L. Vincenzo, “Drift-aware methodology for anomaly detection in smart grid,” *IEEE Access*, vol. 7, pp. 9645-9657, 2019.
- [16]. L. Wen, K. Zhou, Y. Shanlin, and L. Lanlan, “Compression of smart meter big data: A survey,” *Renewable and Sustainable Energy Reviews*, vol. 91, pp. 59-69, 2018.
- [17]. L. Feng, X. Shu, L. Zhang, J. Wu, Z. Jidong, C. Chu, W. Zhenyu and S. Haoyang, “Anomaly detection for electricity consumption in cloud computing: framework, methods, applications, and challenges,” *EURASIP Journal on Wireless Communications and Networking*, vol. 194, 2020.
- [18]. Hela Sfar, B. Amel and B. Raddaoui, “Early anomaly detection in smart home: A causal association rule-based approach,” *Artificial Intelligence in Medicine*, vol. 91, pp. 57-71, September 2018.
- [19]. R. Moghaddass and W. Jianhui, “A hierarchical framework for smart grid anomaly detection using large-scale smart meter data,” *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5820-5830, November 2018.
- [20]. W. Chen, Z. Kaile, S. Yang and C. Wu, “Data quality of electricity consumption data in a smart grid environment,” *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 98-105, 2017.
- [21]. Y. Sook-Chin, T. Wooi-Nee, T. ChiaKwang, G. Ming-Tao and W. KokSheik, “An anomaly detection framework for identifying energy theft and defective meters in smart grids,” *Electrical Power and Energy Systems*, vol. 101, pp. 189-203, 2018.
- [22]. Wei Biao Wu, Maggie X. Cheng and Bei Gou, “A hypothesis testing approach for topology error detection in power grids,” *IEEE Internet of Things Journal*, vol. 3, no. 6, December 2016.
- [23]. Beth-Anne Schuelke-Leech, B. Barry, M. Matteo and B. J. Yurkovich, “Big data issues and opportunities for electric utilities,” *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 937-947, 2015.
- [24]. K. Purna Prakash and Y. V. Pavan Kumar, “Analytical approach to exploring the missing data behavior in smart home energy consumption dataset,” *Journal of Renewable Energy and Environment (JREE)*, vol. 9, no. 2, pp. 1-12, Spring 2022.
- [25]. K. P. Prakash and Y. V. P. Kumar, “Simple and effective descriptive analysis of missing data anomalies in smart home energy consumption readings,” *Journal of Energy Systems*, vol. 5, no. 3, pp. 199-220, 2021.
- [26]. I. Romero-Fiances, L. Andreas, M. Theristis, M. George, J. S. Stein, N. Gustavo, J. de la Casa and G. E. Georghiou, “Impact of duration and missing data on the long-term photovoltaic degradation rate estimation,” *Renewable Energy*, vol. 181, pp. 738-748, 2022.
- [27]. B. Tan, Y. Jun, T. Zhou, X. Zhan, Y. Liu, S. Jiang and L. Chao, “Spatial-temporal adaptive transient stability assessment for power system under missing data,” *Electrical Power and Energy Systems*, vol. 123, pp. 106237, 2020.
- [28]. P. R. Harvey, B. Stephen and G. Stuart, “Classification of AMI residential load profiles in the presence of missing data,” *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1944-1945, July 2016.
- [29]. C. Wang, C. Yu, S. Zhang and L. Tong, “A reconstruction method for missing data in power system measurement based on LSGAN,” *Frontiers in Energy Research*, vol. 9, pp. 1-13, March 2021.
- [30]. I. Izonin, K. Natalia, T. Roman and Z. Khrystyna, “An approach towards missing data recovery within IoT smart system,” *Procedia Computer Science*, vol. 155, pp. 11-18, 2019.
- [31]. C. Genes, I. Esnaola, M. P. Samir, L. F. Ochoa and C. Daniel, “Robust recovery of missing data in electricity distribution systems,” *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4057-4067, July 2019.
- [32]. T. Su, S. Ying, J. Yu, Y. Changxi and Z. Feng, “Nonlinear compensation algorithm for multidimensional temporal data: A missing value imputation for the power grid applications,” *Knowledge-Based Systems*, vol. 215, pp. 106743, 2021.
- [33]. D. Jeong, P. Chiwoo and M. K. Young, “Missing data imputation using mixture factor analysis for building electric load data,” *Applied Energy*, vol. 304, pp. 117655, 2021.

- [34]. S. Ryu, K. Minsoo and H. Kim, "Denoising autoencoder-based missing value imputation for smart meters," *IEEE Access*, vol. 8, pp. 40656-40666, February 2020.
- [35]. J. Ma, J. C. P. Cheng, J. Feifeng, W. Chen, M. Wang and Z. Chong, "A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data," *Energy & Buildings*, vol. 216, pp. 109941, 2020.
- [36]. R. Razavi-Far, M. Farajzadeh-Zanjani, M. Saif and C. Shiladitya, "Correlation clustering imputation for diagnosing attacks and faults with missing power grid data," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1453-1464, March 2020.
- [37]. R. Tawn, J. Browell and D. Iain, "Missing data in wind farm time series: Properties and effect on forecasts," *Electric Power Systems Research*, vol. 189, pp. 106640, 2020.
- [38]. S. Jurado, N. Àngela, F. Mugica and M. Mihail, "Fuzzy inductive reasoning forecasting strategies able to cope with missing data: A smart grid application," *Applied Soft Computing*, vol. 51, pp. 225-238, February 2017.
- [39]. Y. Zhou, S. Lijun, X. Hu and L. Ma, "Clustering and statistical analyses of electricity consumption for university dormitories: A case study from China," *Energy & Buildings*, vol. 245, pp. 110862, 2021.
- [40]. The Tracebase appliance-level power consumption data set, (<http://www.tracebase.org/>)
- [41]. C. Li, "Little's test of missing completely at random," *The Stata Journal*, vol. 13, no. 4, pp. 795-809, 2013.
- [42]. Little's missing completely at random (MCAR) test (https://search.r-project.org/CRAN/refmans/naniar/html/mcar_test.html)
- [43]. A. A. T. Fernandes, D. B. F. Filho, E. Carvalho da Rocha and W. da Silva Nascimento, "Read this paper if you want to learn logistic regression," *Revista de Sociologia e Política*, vol. 28, no. 74, 2020.
- [44]. E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *Journal of Data Analysis and Information Processing*, vol. 7, pp. 190-07, 2019.
- [45]. S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12-18, 2014.
- [46]. P. Mishra, M. P. Chandra, S. Uttam, A. Gupta, S. Chinmoy and K. Amit, "Descriptive statistics and normality tests for statistical data," *Ann Card Anaesth*, vol. 22, no. 1, pp. 67-72, 2019.
- [47]. A. P. King and R. J. Eckersley, *Statistics for Biomedical Engineers and Scientists*, Academic Press, 2019, pp. 147-171.
- [48]. R. C. Aster, B. Borchers and C. H. Thurber, *Parameter Estimation and Inverse Problems*, 3rd ed. Elsevier, 2019, pp. 341-362.
- [49]. K. R. B. Jankowski, F. J. Kevin and F. T. Laura, "The t-test: An influential inferential tool in chaplaincy and other healthcare research," *Journal of Health Care Chaplaincy*, vol. 24, no. 1, pp. 30-39, 2018.
- [50]. Y. Xia, *Progress in Molecular Biology and Translational Science*, Academic Press, vol. 171, pp. 309-491, 2020.
- [51]. G. D. Garson, *Missing values analysis & data imputation*, 2015 Edition, Statistical Publishing Associates, pp. 1-26, 2015.