ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 4015–4023
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.82 Spec Issiue on ICAML-2020

MISSING DATA ANALYSIS FOR ELECTRIC LOAD PREDICTION WITH WHOLE RECORD MISSING

MANDEEP SINGH[1] AND RAMAN MAINI

ABSTRACT. Missing data in the data set used for any data analysis is serious issue for any research. There are various methods that are used for finding the missing data in data set and these are having different conditions for finding missing data from these methods some of methods are for handling the missing data and some of methods are for interpolating or predicting the missing value in data set. In this paper we have been done missing data analysis by using Newton Interpolation methods for finding missing value when we have only one column of data in data set and various values are missing.

1. INTRODUCTION

There are different types of missing data analysis methods and can be categorize further into different forms, [1]. Missing data can be categorized to one of 3 missing data analysis methods, [2] [3]: Data which is missing completely at random (MCAR), data which is missing at random (MAR) and data which is missing not at random (MNAR). Data that are MCAR and MAR are sometimes referred to as ignorable missing data whereas MNAR data is referred to as non-ignorable missing data, [4]. Figure 1 showing the chart for selection of technique for missing data.

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 62N02, 65D05.

Key words and phrases. missing data, MAR, MCAR, MNAR, Newton interpolation.

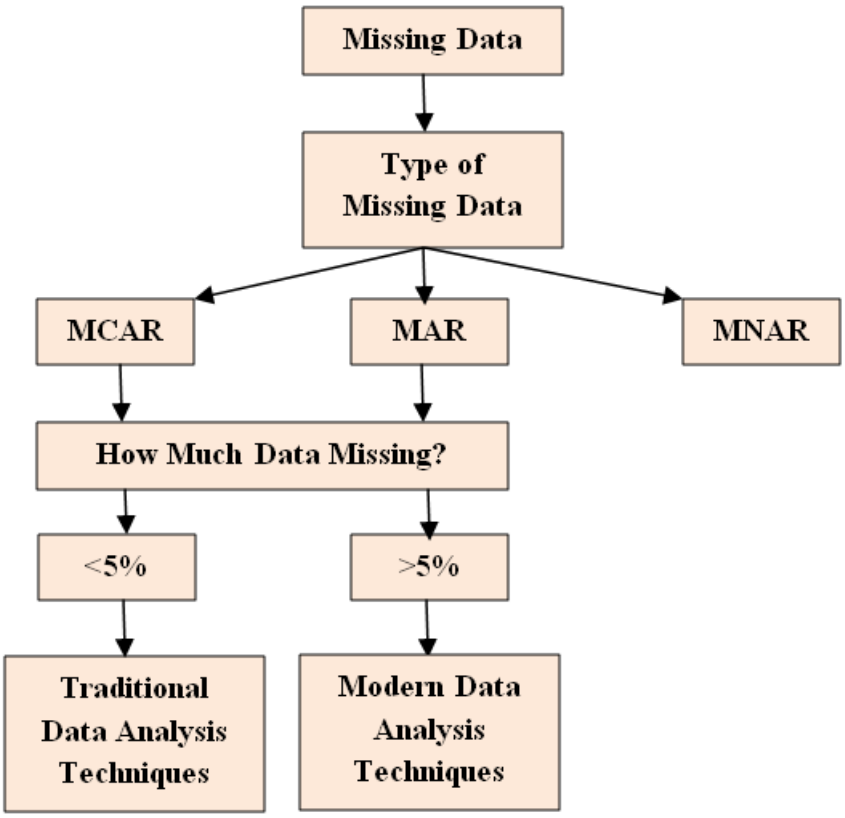

FIGURE 1. Chart for Missing Data techniques

2. MISSING DATA ANALYSIS TECHNIQUES

In general, missing data can be classified into two groups: Traditional Data Analysis and Modern Data Analysis.

2.1. **Traditional Data Analysis Techniques.** The Traditional missing data analysis techniques consist of number of different techniques that can be used to handle missing data. It is a useful tool when a small percentage of the data ($<5\%$) is missing, [5]. The most common traditional techniques are deletion (listwise and pairwise) and single imputation. Single imputation is a process

that involves analyzing the data together with other variables with the intention of finding the most likely value that can be placed in the data.

2.2. **Modern Data Analysis Techniques.** The traditional methods work well for small amounts of missing data. When there is a considerable (>5%) amount of missing data, more sophisticated state of the art techniques and models are required, [5]. Examples, multiple imputations (mi), model-based procedures, machine learning methods.

3. Approach for Finding Missing Values

A clustered approach is used the Parameter r is selected as 5 as observed in the literature this value is sufficient implementing it as follows-

3.1. Newton's backward interpolation method.

(i) Select a clustering parameter r which decides number of values in a cluster. (Assume first $n-1$ values in the data set contain no missing values.)

(ii) Detect the first missing value x placed at position $n+1$ denotes its number in the sequence.

(iii) Propagate backwards by selecting r values of the cluster as $n, n-1, n-2, n-3, \ldots, n-r$

(a) x_n is nth value in the cluster and $n = r$;

(b) x is the missing value number;

(c) h=difference of interval = $n-(n-1) = 1$;

(d) $p = (x - x_n)/h$;

(e) $y_n[\,][\,]$=$n-r; \ldots n-2; n-1; n$ // y_n is an array with one column;

(f) $\nabla y_n = y_n[1][2] - y_n[1][1]; y_n[1][3] - y_n[1][2]; \ldots\ldots$;

(g) $\nabla\, ^n y_n$ = $\nabla\, ^{n-1} y_n$ - $\nabla\, ^{n-1} y_{n-1}$;

(h) Apply Newton backward formula $n_(y^n)$;

(i) $f(x) = y_n + p\nabla y_n + (p(p+1))/2!\nabla\, ^2 y_n + (p(p+1)(p+2))/3!\nabla\, ^3 y_n + \ldots..(p(p+1)(p+2)\ldots.(p+n))/n!\nabla\, ^n y_n$.

(iv) Detect next missing value in the data set.

(v) Repeat step three for all the missing values until all missing values are found. This method is selected because all the values in the beginning are known (assumed for first r values).

3.2. **Newton's forward interpolation method.**

(i) Select a clustering parameter r which decides number of values in a cluster. (Assume first $n-1$ values in the data set contain no missing values.)

(ii) Detect the first missing value x placed at position $n+1$ denotes its number in the sequence.

(iii) Propagate backwards by selecting r values of the cluster as $n, n-1, n-2, n-3, \ldots, n-r$

(a) x_n is nth value in the cluster and $n = r$;

(b) x is the missing value number;

(c) h=difference of interval = $n-(n-1) = 1$;

(d) $p = (x - x_n)/h$;

(e) [ ][ ]= $n-r; \ldots n-2; n-1; n$ // y_n is an array with one column;

(f) $\nabla y_n = y_n[1][2] - y_n[1][1]; y_n[1][3] - y_n[1][2]; \ldots\ldots$;

(g) $\nabla^{\,n} y_n$ = $\nabla^{\,n} y_n$ - $\nabla^{\,n+1} y_{n+1}$;

(h) Apply Newton backward formula $f(x) = y_n + p\nabla y_n + \frac{p(p+1)}{2!}\nabla^{\,2} y_n + \frac{p(p+1)(p+2)}{3!}\nabla^{\,3} y_n + \ldots\ldots \frac{p(p+1)(p+2)\ldots(p+n)}{n!}\nabla^{\,n} y_n$.

(iv) Detect next missing value in the data set.

(v) Repeat step three for all the missing values until all missing values are found. This method is selected because all the values in the beginning are known (assumed for first r values).

3.3. **Newton's central interpolation method.** Newton's central interpolation method is Average of Newton's backward interpolation method and Newton's forward interpolation method.

3.3.1. *Mean method.*

(i) Select a clustering parameter r which decides number of values in a cluster. (Assume first $n-1$ values in the data set contain no missing values.)

(ii) Detect the first missing value x placed at position $n+1$ denotes its number in the sequence.

(iii) Propagate backwards by selecting r values of the cluster as $n, n-1, n-2, n-3, \ldots, n-r$;

(a) x_n is nth value in the cluster and $n = r$;

(b) x is the missing value number;
(c) h=difference of interval = $x = \frac{(\sum n1,n2...n5)}{r}$;

(iv) Detect next missing value in the data set.
(v) Repeat step three for all the missing values until all missing values are found. This method is selected because all the values in the beginning are known (assumed for first r values).

3.3.2. *Median method.*

(i) Select a clustering parameter r which decides number of values in a cluster. (Assume first $n-1$ values in the data set contain no missing values.)
(ii) Detect the first missing value x placed at position $n+1$ denotes its number in the sequence.
(iii) Propagate backwards by selecting r values of the cluster as $n, n-1, n-2, n-3, \ldots, n-r$
(a) x_n is nth value in the cluster and $n = r$;
(b) x is the missing value number;
(c) x= Value at Position $\frac{n+1}{2}$
(iv) Detect next missing value in the data set.
(v) Repeat step three for all the missing values until all missing values are found. This method is selected because all the values in the beginning are known (assumed for first r values).

4. Performance Evaluation of Data

In this paper, we have used Mean, Median, and Newton Interpolation Methods (Newton Forward Interpolation, Newton Backward Interpolation, and Newton Central Interpolation). In tables of results we used N F, N B and N C which represent Newton Forward, Newton Backward and Newton Central Interpolation respectively. In this paper we approach how to use these methods according to application. We use each time n is predicted value and $n-1, n-2, \ldots .n-5$ are previous used values in each approach used to predict the n value in this paper.

4.1. **Prediction Results.** In this section Table 1 shows Prediction result in which Actual Values are the actual values which we already have and used these values with predicted values of various approaches used in this paper. NF, NB, NC,

TABLE 1

Actual Values	Predicted Values				
	NF	NB	NC	Mean	Median
29214	29110	29204	29157	29145	29146
29537	29432	29534	29483	29469	29470
29853	29855	29912	29883	29895	29904
31972	31787	31949	31868	31847	31847
34488	34354	34485	34419	34402	34405

TABLE 2

Actual Errors				
Actual Value – NF	Actual – NB	Actual – NC	Actual - Mean	Actual - Median
-103.65	-9.81	-56.73	-68.29	-67.93
-104.85	-2.72	-53.78	-68.53	-67.45
2.13	59.15	30.64	42.49	51.39
-185.33	-23.33	-104.33	-125.06	-125.17
-134.00	-3.70	-68.85	-86.54	-83.57

Mean and Median shows the predicted values by using our approach to find missing value.

4.2. **Actual Error.** In this section Table 2 shows Actual difference in which Actual-NF shows Actual Value – predicted value (prediction by Newton Forward interpolation approach) and Actual-BF shows Actual Value – predicted value (prediction by Newton Backward interpolation approach) and Actual-NF shows Actual Value – predicted value (prediction by Newton Central interpolation approach) and Actual-Mean shows Actual Value – predicted value (prediction by Mean approach) and Actual-BF shows Actual Value – predicted value (prediction by Median approach).

4.3. **Percentage Errors.** In this section Table 3 shows Percentage Errors in which NF (%) shows percentage of error between actual values and predicated value (prediction by Newton Forward interpolation approach) and NB (%)

TABLE 3

Percentage Errors (%)				
NF(%)	**NB(%)**	**NC(%)**	**Mean(%)**	**Median(%)**
0.35%	0.03%	0.19%	0.23%	0.23%
0.35%	0.01%	0.18%	0.23%	0.23%
0.01%	0.20%	0.10%	0.14%	0.17%
0.58%	0.07%	0.33%	0.39%	0.39%
0.39%	0.01%	0.20%	0.25%	0.24%

shows percentage of error between actual values and predicated value (prediction by Newton Backward interpolation approach) and NC (%) shows percentage of error between actual values and predicated value (prediction by Newton Central interpolation approach) and Mean (%) shows percentage of error between actual values and predicated value (prediction by Mean approach) and Median (%) shows percentage of error between actual values and predicated value (prediction by Median approach).

4.4. **Absolute Errors.** In this section Table 4 shows Absolute difference in which Actual-NF shows Actual Value – predicted value (prediction by Newton Forward interpolation approach) in absolute form and Actual-BF shows Actual Value – predicted value (prediction by Newton Backward interpolation approach) in absolute form and Actual-NF shows Actual Value – predicted value (prediction by Newton Central interpolation approach) in absolute form and Actual-Mean shows Actual Value – predicted value (prediction by Mean approach) in absolute form and Actual-BF shows Actual Value – predicted value (prediction by Median approach) in absolute form.

4.5. **Mean Square Errors.** In this section Table 5 shows Mean Square Errors in which NF shows mean of square error of Prediction (prediction by Newton Forward interpolation approach) in absolute form. NB shows mean of square error of Prediction (prediction by Newton Backward interpolation approach) in absolute form. NC shows mean of square error of Prediction (prediction by Newton Central interpolation approach) in absolute form. Mean shows mean of square error of Prediction (prediction by Mean approach) in absolute form. Median

TABLE 4

Absolute Errors				
Actual - NF	**Actual – NB**	**Actual – NC**	**Actual - Mean**	**Actual - Median**
103.65	9.81	56.73	68.29	67.93
104.85	2.72	53.78	68.53	67.45
2.13	59.15	30.64	42.49	51.39
185.33	23.33	104.33	125.06	125.17
134.00	3.70	68.85	86.54	83.57

TABLE 5

Mean Square Errors				
NF	NB	NC	Mean	Median
14809	**832**	4535	6859	6891

shows mean of square error of Prediction (prediction by Median approach) in absolute form.

5. CONCLUSION

This study introduced different methods available for dealing missing values analysis. The result section shows the analysis results of various approaches used for finding missing value and we found that Newton Backward Interpolation gives best results according to our approach implemented on our problem of finding missing value, highlighted in the Table 5. In future work we can use hybrid approach to find the missing value.

REFERENCES

[1] A. N. BARALDI, C. K. ENDERS: *An introduction to modern missing data analyses*, J. School Psychol., **48**(1) (2010), 5–37.

[2] D. B. RUBIN: *Inference and missing data*, Biometrika, **63**(3) (1976), 581–592.

[3] R. J. A. LITTLE, D. B. RUBIN: *Statistical Analysis with Missing Data*, Wiley, New York, USA, 1987, 381–381.

[4] J. W. GRAHAM: *Missing data analysis: Making it work in the real world*, Annu. Rev. Psychol., **60**(1) (2009), 549–576.

[5] J. DAUWELS, L. GARG, A. EARNEST, L. K. PANG: *Tensor factorization for missing data imputation in medical questionnaires*, Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), (2012), 2109–2112.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
PUNJABI UNIVERSITY PATIALA
PUNJAB, INDIA
Email address: manasdawn@yahoo.com

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
PUNJABI UNIVERSITY PATIALA
$Punjab, India$
Email address: $\texttt{research}_r aman@yahoo.com$