



RESEARCH ARTICLE

REVISED Modified recurrent equation-based cubic spline interpolation for missing data recovery in phasor measurement unit (PMU) [version 3; peer review: 2 approved, 1 not approved]

Shruthi Thangaraj¹, Vik Tor Goh ¹, Timothy Tzen Vun Yap²

¹Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

²Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, 63100, Malaysia

V3 First published: 28 Feb 2022, 11:246
<https://doi.org/10.12688/f1000research.73182.1>
 Second version: 22 Dec 2022, 11:246
<https://doi.org/10.12688/f1000research.73182.2>
 Latest published: 18 Dec 2023, 11:246
<https://doi.org/10.12688/f1000research.73182.3>

Abstract

Background

Smart grid systems require high-quality Phasor Measurement Unit (PMU) data for proper operation, control, and decision-making. Missing PMU data may lead to improper actions or even blackouts. While the conventional cubic interpolation methods based on the solution of a set of linear equations to solve for the cubic spline coefficients have been applied by many researchers for interpolation of missing data, the computational complexity increases non-linearly with increasing data size.








Methods



In this work, a modified recurrent equation-based cubic spline interpolation procedure for recovering missing PMU data is proposed. The recurrent equation-based method makes the computations of spline constants simpler. Using PMU data from the State Load Despatch Center (SLDC) in Madhya Pradesh, India, a comparison of the root mean square error (RMSE) values and time of calculation (ToC) is calculated for both methods.

Results

Open Peer Review

Approval Status   

	1	2	3
version 3 (revision) 18 Dec 2023			 view
version 2 (revision) 22 Dec 2022	 view		  view
version 1 28 Feb 2022	 view	 view	

1. **Mathias Foo** , University of Warwick, Coventry, UK
2. **Shaik Mullapathi Farooq** , K. S. R. M. College of Engineering (UGC-Autonomous), Kadapa, India
3. **Wun She Yap**, Universiti Tunku Abdul Rahman, Kajang, Malaysia

Any reports and responses or comments on the article can be found at the end of the article.

The modified recurrent relation method could retrieve missing values 10 times faster when compared to the conventional cubic interpolation method based on the solution of a set of linear equations. The RMSE values have shown the proposed method is effective even for special cases of missing values (edges, continuous missing values).

Conclusions

The proposed method can retrieve any number of missing values at any location using observed data with a minimal number of calculations.

Keywords

phasor measurement unit, missing data, data recovery, smart grid, interpolation, cubic spline, data quality, data pre-processing



This article is included in the **Artificial Intelligence and Machine Learning** gateway.



This article is included in the **Research Synergy Foundation** gateway.

Corresponding author: Vik Tor Goh (vtgoh@mmu.edu.my)

Author roles: **Thangaraj S:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation; **Goh VT:** Methodology, Supervision, Writing – Review & Editing; **Yap TTV:** Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2023 Thangaraj S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Thangaraj S, Goh VT and Yap TTV. **Modified recurrent equation-based cubic spline interpolation for missing data recovery in phasor measurement unit (PMU) [version 3; peer review: 2 approved, 1 not approved]** F1000Research 2023, 11:246 <https://doi.org/10.12688/f1000research.73182.3>

First published: 28 Feb 2022, 11:246 <https://doi.org/10.12688/f1000research.73182.1>

REVISED Amendments from Version 2

We have addressed the reviewer's concerns, namely the consistency in nomenclature and accuracy of equation numbers. Additionally, we included brief discussions about the results, characteristics of the dataset, and choice of evaluation metrics.

Any further responses from the reviewers can be found at the end of the article

Introduction

The worldwide growing power systems highlight the need for better monitoring and control mechanisms to avoid major blackouts. Smart grids are intelligent systems that facilitate the development of communication, network, and computing technologies, protocols, and standards to integrate power system elements for two-way communication. This time-synchronized high-precision measurement device that is also known as a synchrophasor or Phasor Measurement Unit (PMU), gives clear information on the working of the entire grid. The PMU is used to monitor and control the power grid. It can help in providing real-time measurements by eliminating adverse conditions like blackouts. These combined characteristics of data availability, timeliness, and communication network contribute to the better performance of the PMU system. Although the role, impact,¹ architecture, technology,² applications, functionality, standards, and evolution of PMU (timing, measurement, communication, and data storage) have been released since 1995, the North American Synchro Phasor Initiative (NASPI) has highlighted the importance of data quality.³ Data quality issues, their potential causes, and consequences are elaborated.⁴⁻⁶ Generally, incomplete or missing data might affect the functionality of the entire system.⁷ Hence, a way to handle missing values in PMU is mandatory for the effective functioning of the entire grid system.

With the advent of PMU systems, large datasets are generated and finding missing values using traditional cubic interpolation methods take larger computational time with the increase in data size. In this paper, a modified recurrent equation-based method termed the Alpha Method (AM) for PMU missing data problem is proposed. In this approach, a series of linear equations are solved using the modified recurrent equation to obtain a relationship between points on a spline, which is then used to estimate any missing values on the spline. We compare the proposed method to the more traditional method of solving linear equations, namely using tri-diagonal matrix or termed as the Linear Equations Method (LEM) in this paper. The proposed AM is computationally more efficient and takes less time to process than the LEM. Moreover, in real-time systems when the dataset grows progressively, we show that AM is better than LEM.

Literature review

The need to recover missing values in PMU data is vital to the proper operation of smart grids and the energy infrastructure. Literatures⁵⁻⁷ indicate that missing data in PMU systems can negatively affect the accuracy of decision-making process and additionally, introduce security risks to the infrastructure. To address this problem, missing values have to be recovered and one of the more popular approaches is utilizing matrix completion.⁸⁻¹² Despite that, this approach is still largely theoretical and even so, viable methods utilizing this approach have only been tested on simulated data.

Alternatively, interpolation-based missing data recovery techniques¹³⁻¹⁵ propose the reconstruction of missing values by a spatial interpolation or spatio-temporal interpolation of the values. Some work^{16,17} even suggested advanced approaches utilizing k -nearest neighbors and recurrent relation-based interpolations. However, in interpolation-based techniques, historical data such as channel or time data is needed for more accurate calculations. interpolation. As such, there is a need to design effective data recovery methods to work without the need for historical data processing.³ So, a data-driven recovery technique capable of recovering missing entries with available or observed data is much needed. Moreover, the technique should not become overly complex or require high computational time as the size of the data grows.

Methods

Cubic spline interpolation is a widely used polynomial interpolation method for functions of one variable. Let f be a function from R to R . It is assumed that the value of f is known only at $x_1 \leq x_2 \leq x_i \dots \leq x_n$ and let $f(x_i) = a_i$. Piecewise cubic spline interpolation is the problem of finding the b_i , c_i and d_i coefficients of the cubic polynomials SF_i for $0 \leq i \leq n - 1$ written in the form:

$$SF_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (1)$$

Where x can take any value between x_i and x_{i+1} . That is,

$$SF_i(x_i) = a_i \tag{1a}$$

Let the first-order derivative of [equation \(1\)](#) be:

$$SF'_i(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \tag{2}$$

The first-order derivative at x_i for values of $1 \leq i \leq n - 1$ will be

$$SF'_i(x_i) = b_i \tag{2a}$$

And the second-order derivative be:

$$SF''_i(x) = 2c_i + 6d_i(x - x_i) \tag{3}$$

The second-order derivative at x_i for values of $1 \leq i \leq n - 1$ will be:

$$SF''_i(x_i) = 2c_i \tag{3a}$$

For a smooth fit between the adjacent pieces, the cubic spline interpolation requires that the following conditions hold:

1. The cubic functions should intersect at the points left and right, for $i = 0$ to $n - 1$

$$SF_i(x_{i+1}) = SF_{i+1}(x_i) = a_{i+1} \tag{4}$$

2. For each cubic function to join smoothly with its neighbors, the splines should have continuous first and second derivatives at the data points $i = 1, \dots, n - 1$:

$$SF'_i(x_{i+1}) = SF'_{i+1}(x_i) = b_{i+1} \tag{5}$$

$$SF''_i(x_{i+1}) = SF''_{i+1}(x_i) = 2.c_{i+1} \tag{6}$$

If $h_i = x_{i+1} - x_i$ and if h_i is equal for all i values, following Revesz,¹⁷ the relation between coefficients a_i and c_i can be resolved:

$$c_{i-1} + 4c_i + c_{i+1} = \frac{3}{h_i^2}(a_{i-1} - 2a_i + a_{i+1}) \tag{7}$$

$$b_i = (a_{i+1} - a_i) \frac{1}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i \tag{8}$$

$$d_i = \frac{1}{3.h_i}(c_{i+1} - c_i) \tag{9}$$

[Equation \(7\)](#) represents a system of linear equations for the unknowns c_i for $0 \leq i \leq n$. As the values of a_i are known, the value of c_i can be found by solving the tri-diagonal matrix-vector equation $Ax = B$. While there are $n+1$ numbers of c_i constants, [equation \(7\)](#) yields only $(n-2)$ equations. Based on the nature or type of spline assumed two more equations representing the boundary conditions of the spline. In general, two types of splines may be considered: natural cubic spline and clamped cubic spline.

For natural cubic spline interpolation, the following boundary conditions are assumed: $c_0 = c_n = 0.0$. That is, the second derivatives of the splines at the endpoints are assumed to be zero. Based on [equation \(7\)](#), a system of $(N+1)$ linear equations of $(N+1)$ variables can be formulated as:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}, x = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}, \text{ and } B = \begin{bmatrix} 0 \\ \frac{3}{h^2}(a_0 - 2a_1 + a_2) \\ \vdots \\ \frac{3}{h^2}(a_{n-2} - 2a_{n-1} + a_n) \\ 0 \end{bmatrix} \quad (10)$$

For clamped cubic spline interpolation the following boundary conditions are assumed: $b_0 = f'(x_0)$ and $b_n = f'(x_n)$, where the derivatives $f'(x_0)$ and $f'(x_n)$, are known constants. Thus, based on the boundary conditions assumed both natural and cubic splines result in $n+1$ system of linear equations. The resulting system of $n+1$ linear equations can be used to get unique solutions by any of the standard methods for solving a system of linear equations.

Once the values of c_i are found, the b_i and d_i values can be obtained using equations (8) and (9) respectively. Similarly, under clamped spline interpolation,

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 2 \end{bmatrix}, x = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} \text{ and } B = \begin{bmatrix} \frac{3}{h^2}(a_1 - a_0) - \frac{3}{h}f'(x_0) \\ \frac{3}{h^2}(a_0 - 2a_1 + a_2) \\ \vdots \\ \frac{3}{h^2}(a_{n-2} - 2a_{n-1} + a_n) \\ \frac{3}{h}f'(x_0) - \frac{3}{h^2}(a_n - a_{n-1}) \end{bmatrix} \quad (11)$$

Recurrent equation-based solution

Revesz,¹⁷ chose boundary conditions that need to solve the tri-diagonal system given in equation (7) where x_i rational variables e_i rational constants, r is a non-zero rational constant and A is:

$$A = \begin{bmatrix} r & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \text{ and } b = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix} \quad (12)$$

The first row of the new matrix in equation (12) is shown to be equivalent to the first row of the clamped b matrix e_1 is

$$e_1 = \frac{3r}{2h} \left(\frac{a_1 - a_0}{h} - f'(x_0) \right) + \left(1 \frac{r}{2} \right) \tilde{c}_1 \quad (13)$$

where, \tilde{c}_1 is an estimate of c_1 and $r = 2 + \sqrt{3} \approx 3.732$.¹⁷

The chosen boundary conditions are such that the first row of the new matrix was the same as that of clamped cubic spline and while that of the last row was that of the natural cubic spline fixing the value of c_n as 0. Using equation (12), the relationships between successive spline points can be obtained as:

$$x_i + \frac{x_{i+1}}{r} = \sum_{0 \leq k \leq (i-1)} (-1)^k \frac{e_{i-k}}{r^{k+1}} \quad (14)$$

Let α_0, α_i for $1 < i \leq n - 1$ and α_n , respectively be:

$$\alpha_0 = 0$$

$$\alpha_i = \frac{e_i - \alpha_{i-1}}{r} = \sum_{0 \leq k \leq (i-1)} (-1)^k \frac{e_{i-k}}{r^{k+1}} \quad (15)$$

$$\alpha_n = e_n$$

Based on the above, the closed form of solution for x_i can be given as:

$$x_i = \sum_{0 \leq k \leq (n-i)} \left(\frac{-1}{r}\right)^{-k} \alpha_{i+k} \quad (16)$$

The above equation (16) solves x_i no matter exactly what the initial values for e_i . This leads to a faster evaluation of the cubic spline than solving a tri-diagonal system. The major advantage of the method is when new measurements are added to the system. While conventional tri-diagonal matrix-based algorithm requires a complete redo of the entire computation, equation (16) leads to a faster update for each $i \leq n$ only with the addition of the term:

$$\left(\frac{-1}{r}\right)^{n+1-i} \alpha_{n+i} \quad (17)$$

and $x_{n+1} = \alpha_{n+1}$. Similarly, α_i constants can be updated by adding a single term e_{n+1}

The system of linear equations given in equation (7), in general, is solved by the standard solution of linear equations in the matrix form $Ax = b$. Alternatively, it could be solved for n variables by the recurrence relations given equations (16) and (17). The two methods, the first using the tri-diagonal matrix-based solution for the spline coefficients is termed the Linear Equations Method (LEM) and the second one using recurrence relations is termed the Alpha Method (AM). The algorithmic procedure for LEM and AM are given below.

Algorithmic procedure for regular tridiagonal matrix-based Linear Equation Method (LEM)

Step 1: Given the initial vector with missing values, separate them into two sets of vectors, the observed values vector R_{obs} and the missing values vector R_{Miss} , having sizes of NO and NM , respectively, such that $NO+NM=N$.

Step 2: R_{obs} vector at x_i values of the $(NO-1)$ splines shall be the a_i coefficient vector.

Step 3: Using a_i , generate the RHS vector B given in equation (11).

Step 4: Generate a square coefficient matrix A as given in equation (11)

Step 5: Solve for the c_i vector is given in (11), using the relation $Ax = B$

Step 6: Applying c_i in equations (8) and (9), compute the b_i and d_i coefficient vectors for $n-2$ points of the R_{obs} .

Step 7: Using the values of a_i, b_i, c_i and d_i , missing values can be found by the equation (1) re-written as:

$$SF_i(x) = (a_i - b_i x_i + c_i x_i^2 - d_i x_i^3) + (b_i - 2c_i x_i + 3d_i x_i^2)x + (c_i - 3d_i x_i)x^2 \quad (18)$$

Where x represents the missing positions, between x_i and x_{i+1} of spline i .

Algorithmic procedure for recurrent equation-based Alpha Method (AM)

Step 1: Given the initial vector with missing values, separate them into two sets of vectors, the observed values vector R_{obs} and the missing values vector R_{Miss} , having sizes of NO and NM , respectively, such that $NO+NM=N$.

Step 2: The R_{obs} vector at x_i values of the $(NO-1)$ splines is the a_i coefficient vector.

Step 3: Using a_i , generate the RHS vector B given in [equation \(11\)](#).

Step 4: Set $\alpha_0 = 0$ and $\alpha_n = e_n$, calculate the alpha vector using the relation.

$$\alpha_i = \frac{e_i - \alpha_{i-1}}{r} = \sum_{0 \leq k \leq (i-1)} (-1)^k \frac{e_{i-k}}{r^{k+1}} \text{ for } i \text{ values ranging from } 1 \text{ to } NO-1$$

Step 5: Set $x_n = \alpha_n$ and solve for c_i values using the relation.

$$c_i = \sum_{0 \leq k \leq (n-i)} \left(\frac{-1}{r}\right)^k \alpha_{i+k}$$

Step 6: Applying c_i in [equations \(8\)](#) and [\(9\)](#), compute the b_i and d_i coefficient vectors for $n-2$ points of the R_{obs} .

Step 7: Using the values of a_i, b_i, c_i and d_i , missing values can also be found using [equation \(18\)](#), re-written here again for convenience:

$$SF_i(x) = (a_i - b_i x_i + c_i x_i^2 - d_i x_i^3) + (b_i - 2c_i x_i + 3d_i x_i^2)x + (c_i - 3d_i x_i)x^2 \quad (18)$$

Where x represents the missing positions, between x_i and x_{i+1} of spline i .

The modifications are as follows: In AM, rather than computing E , alpha vectors and c_i coefficients for the full range of $NO-1$ data points only the RHS, E vector, was calculated for the full range of $NO-1$ data points, while alpha vector and c_i were calculated only for i and $i + 1$ data elements, where i is the missing data element. For the imputation of i the element, only the E_i vector for all $NO-1$ data points, α_i vector and c_i vectors for i and $i + 1$ and b_i and d_i coefficients were essential for the calculation i^{th} missing element and its imputation.

In addition, using the AM, an effective procedure was demonstrated for the computation of the following cases: (i) missing first and the last element of the data vector, (ii) missing multiple data points at the beginning and the end, and (iii) missing multiple elements anywhere in the data vector. That is in [equation \(18\)](#), when the current values of $A [i]$ are replaced either with $A [N-1]$ or $A [i-1]$ based on the position of missing edge values or continuous values the Time of Calculation (ToC) and Root Mean Squared Error (RMSE) values have improved significantly.

The formula for RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{Predicted} - \text{Actual})^2}{N}}$$

We have used RMSE and ToC as evaluation metrics to measure the effectiveness and efficiency of the proposed method because most literature used the same.

Results and discussion

A comparison between LEM and AM is shown here for the imputation of one-min real PMU system data having a size of 1490 data points for each of the 25 heterogeneous variables obtained from five different PMUs. Since our data does not have any missing values, we artificially introduced the missing values, of 10%, 20%, 30% in random.

A sample of one-minute PMU data for five PMUs' was used in the study.¹⁸ One minute of PMU data with 10%, 20%, and 30% missing data for five PMUs were evaluated.

When AM was employed, the average RMSE values were 0.83, 1.47, and 2.16 for 10%, 20%, and 30% of missing PMU data, respectively. This can be seen in [Figure 1](#). Moreover, for the same performance, AM showed significant improvements in its ToC as shown in [Figure 2](#). The average ToCs for AM were 1.35, 1.41, and 1.23s when recovering 10%, 20%, and 30% of its missing data.

By comparison, LEM had ToC values of 18.83, 16.02, 16.58s for 10%, 20%, and 30% of its missing data, respectively. The proposed method reduced the ToC by a factor of approximately 10 times. LEM had higher ToC values because it needed to solve the entire set of linear equations every time it needed to find the b_i, c_i , and d_i coefficients. On the other hand, AM only needed to calculate these coefficients at two successive points of i and $i+1$.

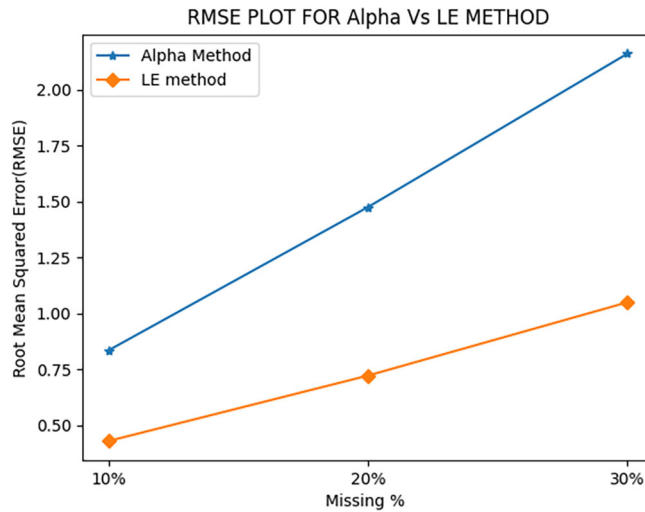


Figure 1. Comparison of Root Mean Squared Error (RMSE) values.

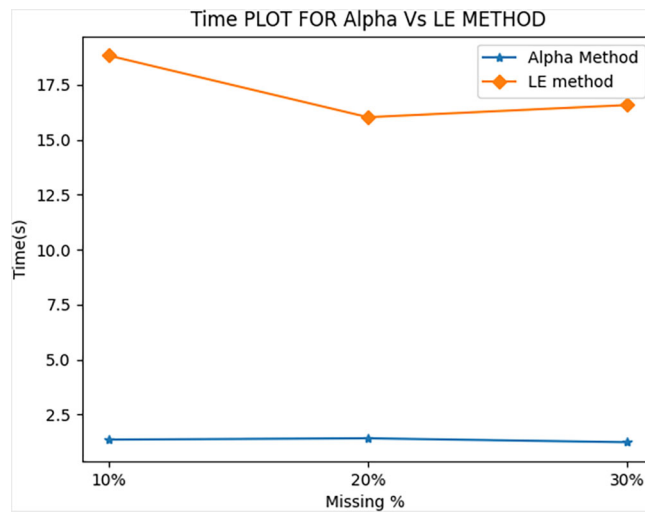


Figure 2. Comparison of Time of Calculation (ToC).

Conclusions

In this study, AM was compared with LEM. However, because of the proliferation of the data, there is a need for customization of this technique to handle a high volume of data to reduce computational time and power. In the proposed method, the approaches demonstrated a reduced computational effort and time of calculation for solving the coefficient vectors. This study has made the following contributions: (i) the recurrent relation-based AM has been effectively employed in the imputation of PMU data and its advantages are demonstrated as an effective and efficient alternative to the conventional technique, and (ii) an effective procedure for handling missing values in special cases (edge, continuous values) is shown, which has not been addressed clearly in other methods. The proposed method has proven effective, and it only requires 10% effort in comparison to the LEM. Future research will focus on the application of the modified recurrent method in the analysis of real-time or stream PMU data.

Data availability

Underlying data

Harvard Dataverse: Underlying data for ‘Modified recurrent equation-based cubic spline interpolation for missing data recovery in phasor measurement unit (PMU)’, ‘PMU data’, <https://doi.org/10.7910/DVN/Y2LLJJ>.¹⁸

This project contains the following underlying data:

- Data file: pmu1-1m-10.tab – One minute of data from PMU1 with 10% missing data
- Data file: pmu1-1m-20.tab – One minute of data from PMU1 with 20% missing data
- Data file: pmu1-1m-30.tab – One minute of data from PMU1 with 30% missing data
- Data file: pmu2-1m-10.tab – One minute of data from PMU2 with 10% missing data
- Data file: pmu2-1m-20.tab – One minute of data from PMU2 with 20% missing data
- Data file: pmu2-1m-30.tab – One minute of data from PMU2 with 30% missing data
- Data file: pmu3-1m-10.tab – One minute of data from PMU3 with 10% missing data
- Data file: pmu3-1m-20.tab – One minute of data from PMU3 with 20% missing data
- Data file: pmu3-1m-30.tab – One minute of data from PMU3 with 30% missing data
- Data file: pmu4-1m-10.tab – One minute of data from PMU4 with 10% missing data
- Data file: pmu4-1m-20.tab – One minute of data from PMU4 with 20% missing data
- Data file: pmu4-1m-30.tab – One minute of data from PMU4 with 30% missing data
- Data file: pmu5-1m-10.tab – One minute of data from PMU5 with 10% missing data
- Data file: pmu5-1m-20.tab – One minute of data from PMU5 with 20% missing data
- Data file: pmu5-1m-30.tab – One minute of data from PMU5 with 30% missing data
- README.txt

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

The dataset presented in the work was obtained as real-world data from a regional Electricity authority in India. However, additional information such as the data source, the acquisition procedure, and the significance of the systemic variables are not detailed at this stage of algorithm development as the goal of this preliminary work is to demonstrate the efficacy of the proposed missing data recovery algorithm.

References

1. Phadke AG, Bi T: **Phasor measurement units, WAMS, and their applications in protection and control of power systems.** *J. Mod. Power Syst. Clean Energy.* 2018; 6(4): 619–629.
[Publisher Full Text](#)
2. Usman MU, Faruque MO: **Applications of synchrophasor technologies in power systems.** *J. Mod. Power Syst. Clean Energy.* 2019; 7(2): 211–226.
[Publisher Full Text](#)
3. Amidan B, et al.: *Data Mining Techniques and Tools for Synchrophasor Data.* North American SynchroPhasor Initiative (NASPI); 2019, January; 45.
4. Miller LE, et al.: **PMU Data Quality: A Framework for the Attributes of PMU Data Quality and a Methodology for Examining Data Quality Impacts to Synchrophasor Applications.** 2017; no. March: pp. 1–77.
5. Huang C, et al.: **Data quality issues for synchrophasor applications Part I: a review.** *J. Mod. Power Syst. Clean Energy.* 2016; 4(3): 342–352.
[Publisher Full Text](#)
6. Huang C, et al.: **Data quality issues for synchrophasor applications Part II: problem formulation and potential solutions.** *J. Mod. Power Syst. Clean Energy.* 2016;

- 4(3): 353–361.
[Publisher Full Text](#)
7. Fang X, *et al.*: **PMU Data Quality: A Framework for the Attributes of PMU Data Quality and a Methodology for Examining Data Quality Impacts to Synchrophasor Applications.** *IEEE Trans. Power Syst.* 2017; **7**(1): 1–6.
 8. Genes C, Esnaola I, Perlaza SM, *et al.*: **Recovering missing Data via matrix completion in electricity distribution systems.** *IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC, 2016-Augus (July 2016).* 2016.
 9. Gao P, Wang M, Ghiocel SG, *et al.*: **Missing Data Recovery by Exploiting Low-Dimensionality in Power System Synchrophasor Measurements.** *IEEE Trans. Power Syst.* 2016; **31**(2): 1006–1013.
[Publisher Full Text](#)
 10. Cai JF, Candès EJ, Shen Z: **A singular value thresholding algorithm for matrix completion.** *SIAM J. Optim.* 2010; **20**(4): 1956–1982.
[Publisher Full Text](#)
 11. Genes C, Esnaola II, Perlaza SM, *et al.*: **Recovering missing data via matrix completion in electricity distribution systems.** *IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC, 2016-August, 1–6.* 2016.
 12. Hastie T, Mazu Missing Dataer R, Lee JD, *et al.*: **Matrix completion and low-rank SVD via fast alternating least squares.** *J. Mach. Learn. Res.* 2015; **16**: 3367–3402.
[PubMed Abstract](#)
 13. Gräler B, Pebesma E, Heuvelink G: **Spatio-temporal interpolation using gstat.** *R Journal.* 2016; **8**(1): 204–218.
[Publisher Full Text](#)
 14. Cheng S, Lu F: **A two-step method for missing spatio-temporal Data reconstruction.** *ISPRS Int. J. Geo Inf.* 2017; **6**(7).
[Publisher Full Text](#)
 15. Deng M, Fan Z, Liu Q, *et al.*: **A Hybrid Method for Interpolating Missing Data in Heterogeneous Spatio-Temporal Datasets.** *ISPRS Int. J. Geo Inf.* 2016; **5**(2).
[Publisher Full Text](#)
 16. Yang Z, Liu H, Bi T, *et al.*: **A PMU data recovering method based on preferred selection strategy.** *Glob. Energy Interconnect.* 2018; **1**(1): 63–69.
 17. Revesz PZ: **A recurrence equation-based solution for the cubic spline interpolation problem.** *International Journal of Mathematical Models and Methods in Applied Sciences.* 2015; **9**(16): 446–452.
 18. Thangaraj S, Goh VT, Yap TTV: **PMU Data.** 2021. Harvard Dataverse.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:   

Version 3

Reviewer Report 22 December 2023

<https://doi.org/10.5256/f1000research.160163.r231067>

© 2023 Yap W. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Wun She Yap

Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang, Malaysia

The authors have addressed the concerns raised in my peer review report appropriately.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial intelligence, cryptography, information security

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 27 November 2023

<https://doi.org/10.5256/f1000research.142368.r220241>

© 2023 Yap W. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Wun She Yap

¹ Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang, Malaysia

² Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang, Malaysia

The paper proposed a method to perform missing data recovery on PMU data. It uses a modified approach based on an established technique i.e. recurrent equation-based cubic spline interpolation. Although the paper can be difficult to follow in some sections especially the Methodology, overall, the method and its results do show some promise and is worth further investigation.

This paper can be considered for publication if the suggestions below are addressed.

Literature can be improved with a discussion about of some current missing data recovery techniques, especially the technique that the proposed method is based on, that is the recurrent equation-based method and tri-diagonal matrix-based conventional cubic spline interpolation. Those techniques are not discussed in the literature review despite their apparent importance to this paper.

Although not required in this paper due to the need for conciseness and perhaps page limits, it is recommended that the authors use some visual aids/figures/plots to help explain the notion of cubic spline in future manuscripts.

In Page 5, the section heading called "Recurrence equation-based solution" should be "Recurrent equation-based solution" for consistency with the rest of the text. Similarly, please ensure consistency throughout the paper. There are inconsistent terms such as LEM and LE (in Introduction), AM and AM Method (using method is redundant), etc.

Please check the numbering for the equations and the references/citations to them. For example, in Page 5, the authors state "The first row of the new matrix in (6) is shown..." would imply that Equation 6 is a matrix but Equation 6 is just a normal formula.

The dataset is lacking context but understandable due to the confidentiality of the data. It would help if more general characteristics of the dataset are given or an anonymised sampling is shown in a table. That would add depth to the discussion.

The discussion provides valuable insights into the comparison of the two methods, but additional details and context in certain areas could enhance the clarity and completeness of the findings. For example, the use of RMSE as an evaluation metric is common in imputation tasks, but it would be helpful to know if other metrics were considered or if there are specific reasons for choosing RMSE. Similarly, why was ToC used as a metric?

The comparison results that are provided in terms of RMSE and ToC seem to indicate that while Alpha Method has higher RMSE values compared to LE method, AM performs better in terms of time. Providing some insights in the discussion section into why this improvement occurred (e.g., the nature of the algorithm, computational complexity) would enhance the discussion.

Can the proposed method be used for other datasets with missing data or is it optimised specially for PMU data? A brief explanation to clarify the impact of PMU data would further improve the discussion.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial intelligence, cryptography, information security

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Dec 2023

Vik Tor Goh

The literature review has been amended to improve its readability and conciseness.

We will include such visual aids in our future publications. Thank you for the suggestion.

The necessary changes have been made to ensure consistency in the nomenclatures and redundancies have been eliminated.

The numbering for the equations has been amended accordingly.

As pointed out, it will not be possible to share the full dataset due to confidentiality reasons. We briefly explained some characteristic in the Results and Discussion section and added some additional statements of the dataset in the Data Availability section. Furthermore, the dataset can be readily downloaded from the public repository and examined.

From our literature review, we determined that most literature use RMSE as an evaluation metric. As such, we adopt the same metric to ensure that our work can be fairly compared with other methods. We have added further clarification about this in the manuscript.

LEM had higher ToC values because it needed to solve the entire set of linear equations every time it needed to find the b_i , c_i , and d_i coefficients. On the other hand, AM only needed to calculate these coefficients at two successive points of i and $i+1$. A brief discussion has been added in the Results and Discussion section explaining this.

The proposed data recovery method is dataset-neutral and can be used in any big data system. However, in our work, we utilized PMU data as a case study because PMU data was “big” and most importantly, it was readily available to us. For our future work, we will apply the proposed method on other datasets to examine its viability.

Competing Interests: No competing interests were disclosed.

Reviewer Report 03 January 2023

<https://doi.org/10.5256/f1000research.142368.r158516>

© 2023 Foo M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Mathias Foo 

¹ School of Engineering, University of Warwick, Coventry, UK

² School of Engineering, University of Warwick, Coventry, UK

The authors have addressed all my comments and the new manuscript reads better than before.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Dynamical system modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 July 2022

<https://doi.org/10.5256/f1000research.76818.r139819>

© 2022 Farooq S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Shaik Mullapathi Farooq 

¹ Department of Computer Science and Engineering, K. S. R. M. College of Engineering (UGC-Autonomous), Kadapa, Andhra Pradesh, India

² Department of Computer Science and Engineering, K. S. R. M. College of Engineering (UGC-Autonomous), Kadapa, Andhra Pradesh, India

The manuscript proposes recurrent relation based alpha method to interpolate missing PMU data. Further, the authors try to prove that the proposed method reduces computational complexity.

However, the comments are as follows,

1. The Implementation of the proposed method is clearly missing (Hardware or software details used) in the manuscript which does not assist reproducing the results.
2. Most of the manuscript is dedicated for theoretical discussion about the proposed method. But a comparison between the existing methods with the proposed method is missing.
3. Add nomenclature that improves the readability of the manuscript.
4. Only data set of PMU values are presented (PMU Data Harvard Data verse) instead need to add discussion about the details of PMU Data.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cyber security in smart grid communication network and VANET.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 16 Dec 2022

Vik Tor Goh

The purpose of this preliminary paper is to introduce our work in missing data recovery using cubic spline interpolation, namely the mathematical foundation and algorithmic logic. These details have been presented and explained accordingly in the paper. Additionally, the data used is also available for download by interested parties. We aim to publish a more detailed paper soon which will contain more information such as those suggested by the reviewer. Thank you for the suggestion.

As stated earlier, the purpose of this preliminary paper is to introduce our proposed method, hence the emphasis on theoretical discussion. However, we have made an initial comparison with an existing method, namely the Linear Equation Method. This can be seen in the Results and discussion section.

The nomenclature is improved wherever possible to improve the readability of the manuscript.

The dataset presented in the work was obtained from a regional Electricity Authority in India. It was obtained for use as realistic data and brief details of the PMU data is now included. However, additional information such as the data source, the acquisition process, and the physical significance of the systemic variables are not detailed at this stage of algorithmic development as the main idea is only to demonstrate the efficacy of the missing data imputation algorithm. Nonetheless, we take note of this suggestion for our next submission. Thank you.

Competing Interests: No competing interests were disclosed.

Reviewer Report 10 March 2022

<https://doi.org/10.5256/f1000research.76818.r125681>

© 2022 Foo M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Mathias Foo** 

¹ School of Engineering, University of Warwick, Coventry, UK

² School of Engineering, University of Warwick, Coventry, UK

In general, there is promising aspect of the proposed method but it has to be conveyed in a clearer manner. Here are my comments.

1. In Introduction section, the authors state that the comparison will be made with LEM. Can the author explain why specifically LEM is compared? Is that the current state-of-the-art method?
2. In Literature Review section, NASPI is mentioned but without proper definition of the acronym.
3. If my understanding is right, Equation (10) is a systems of linear equation of (7). Then, why does the h value in the B matrix have an exponent of 1 instead of 2 as of Equation (7)?
4. Statement above Equation (11): Unless I'm mistaken, there is no d coefficient to be solved from either equations (5) or (6).
5. Equation (11): Like Equation (10), can the authors clarify why the exponent of 1 is used for h?
6. Equation (13): Why is r taking this value? A bit more explanation would be helpful.
7. Equation (16): Why is there a 'for' in the equation?
8. Step 3 of LE method: There is no vector E in Equation (11).
9. Step 3 of AM method: There is no vector E in Equation (11).
10. Step 4 of AM method: There is no alpha term in Equation (11).
11. Results and Discussions section: Can the author explicitly write down the equation for RMSE?

Also, I am quite surprised with the huge difference in terms of RMSE between the two methods even for the case of 10% missing data considering the same equation (18) is used for both algorithms. The difference in ToC is understandable, but the vast difference in RMSE is a bit out of my expectation. Could the author briefly comment on the plausible reason for this huge difference in the RMSE value despite both algorithm using equation

(18).

12. Overall comment: The mathematical derivation is not easy to follow and there are potential mistakes in citing the equations, which makes it even harder to follow. Thus, it is difficult to ascertain whether the results can be reproduced.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Dynamical system modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Dec 2022

Vik Tor Goh

The idea of cubic spline is the development of a series of unique cubic polynomials that are fitted between the data points. Based on four continuity relations between points in the spline, the relationships between the spline coefficients shall result in a system of unique $n \times n$ linear equations in the matrix form $Ax = B$. The solution of this unique system of linear equations results in the values of constants at each spline point. Whenever some changes occur in any one of the splines, the system of linear equations must be solved for every specific change to fit the spline. The tri-diagonal method of solving a system of linear equations was employed in this study for comparison. Hence and from the reference, we made such a comparison to linear equation method.

The word is already introduced in the introduction section as North American Synchro

Phasor Initiative (NASPI).

The exponent for h in Equation (10) and (11) should be 2 instead of 1. We have made the corrections accordingly. Thank you.

As the reviewer correctly noted, the d coefficient is not solved in Equations (5) or (6). It is instead solved using Equation (9). We have made the amendments accordingly.

The r -value in Equation (13) is a non-zero rational constant; the value is adapted from our reference work [17]. We have added this citation in the text.

The 'for' should not be in Equation (16). It has been corrected.

Step 3 and Step 5 of the LE method has been corrected as vector B instead of E .

Reference to Equation (11) in Step 4 of the AM method has been removed.

The equation for RMSE has been added in the text.

Upon inspection, we found that the variables b , c , and d of the cubic spline is found to be similar in both methods as the reviewer correctly predicted. We re-examined our results and have determined that due to an oversight, an error occurred in the final calculations of imputation values in the LE method. Instead of the coordinate numbers, the spline values at the coordinates were used for the calculation of missing values. As such, we have made the corrections to the results and discussions, as well as the plots in Figure 1 and Figure 2.

All corrections and suggestions given were incorporated and the revised version of the paper is written. We really appreciate the time spent by the reviewer for the useful suggestions and corrections made.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research