A Two-Stage Inference Method Based on Graph Neural Network for Wind Farm SCADA Data

Zhanhong Ye, Fan Wu(✉), Cong Zhang, Wenhao Fan, Bihua Tang, and Yuanan Liu

Beijing University of Posts and Telecommunications, Beijing 100876, China
wufanwww@bupt.edu.cn

Abstract. Wind power generation is a representative of high-quality new energy. Real-time monitoring and accurate prediction of wind turbines are critical to ensure their stable operation. Due to sensor failures, network congestion, and communication errors, wind turbine monitoring data are often accompanied by data losses which affects the performance of the wind power prediction model. To address the challenge, we propose a two-stage method for inferring missing values in wind power data. First, the missing value supplement and selection of variables with high similarity in changes are applied, and the top-k nearest neighbors are employed to construct coarse-grained estimation. Second, we proposed a multi-view graph learning framework to capture the latent representation of wind power data from three views. The missing values will be inferred based on these latent representations. Finally, experiments with real world data demonstrate that our method has better inference accuracy than traditional and deep learning inference methods.

Keywords: wind power data · continuous missing · Graph neural networks · spatio-temporal

1 Introduction

Nowadays, wind energy has become one of the main renewable energy sources accommodated by smart grids. In 2022, 77.6 GW of new wind power capacity was connected to the grid worldwide, with a total installed capacity of 906 GW, representing a 9% increase compared to 2021 [1]. Within this context, accurately predicting wind turbine(WT) power based on existing monitoring data and making timely regulations is crucial for alleviating the pressure on real-time smart grid dispatch. However, due to sensor failures, network congestion, and communication errors, there are varying degree of data loss in the WT Supervisory Control and Data Acquisition (SCADA) system. These data losses pose various obstacles to the operation and maintenance of wind farms, such as fault diagnosis [2], status monitoring, and power generation prediction [3]. On the one

X. Li et al. (Eds.): SmartGift 2024, LNICST 600, pp. 100–111, 2025.
https://doi.org/10.1007/978-3-031-78806-2_6

hand, missing data cannot be directly processed by many data-driven prediction models, which significantly reduces the model's expected performance. On the other hand, the lack of some important monitoring data can directly affect the status adjustment of WTs, potentially leading to more severe failures. Therefore, ensuring the integrity and accuracy of SCADA data has great significance in intelligent regulation of wind power systems.

Existing methods for handling missing data can be divided into two categories: deletion and imputation. In smart grids, SCADA systems can contain values for many attributes, simply discarding entire incomplete records due to a few missing values may not be ideal. These shortcomings limit the popularity of deletion methods in practice. Imputation refers to the application of specific algorithms to fill in missing values and obtain a complete dataset. Compared to deletion methods, imputation methods are widely used but differ in applicability and accuracy.

Imputation methods can be roughly divided into two categories: classical statistical methods and data-driven methods. Interpolation methods [4], nearest neighbors (KNN) [5], k-means, autoregressive integrated moving average models, etc., are effective methods in statistics for recovering missing values. These methods are often based on a series of strict assumptions, such as linearity, normality, and independence and identical distribution. Nevertheless, high-dimensional data frequently deviate from these assumptions owing to their intricate nonlinear connections, heteroscedasticity, and correlations. The disadvantages of statistical methods which can not accurately describe the complex structures and patterns in high dimensional wind power data, leading to the efficiency of data imputation significantly decreasing when missing rate increases. As a new powerful method, data-driven methods often achieve better results when applied to the imputation of missing data in wind farms. A new LSTM model has been proposed to interpolate missing data in multivariate time series [6]. A new attention-based architecture has been introduced to reconstruct observations regarding corresponding sensors and their adjacent nodes by leveraging a spatiotemporal propagation architecture consistent with the imputation task [7]. A denoising autoencoder based on a spatiotemporal graph neural network has been developed to perform inference on spatiotemporal sequences of wind power data [8]. Compared with classical statistical methods, data-driven based method fully learned the complex data patterns of different WTs, resulting in higher accuracy and wider application. But these methods consider the physical distance between graph nodes as adjacency matrix which is static, while the wind data between adjacent wind farms exhibits dynamic relationships. Additionally, these methods ignore the feature correlation between different attributes of data in WTs monitoring. Furthermore, missing data are often set to zero as the input of the neural network, which can lead to slow network convergence and affect the extraction of spatiotemporal features when there are a large of missing values.

Given the variability in missing data patterns across different scenarios, the imputation performance of the aforementioned data-driven methods may be different. In many studies [11,12], simulation experiments were conducted by ran-

domly removing parts of the data and generating inferred data to validate the performance of imputation methods. According to the form of missing wind power data, its missing types can be divided into continuous missing and dispersive missing [10,11], as shown in Fig. 1. The former is due to sensor failure of WTs, which usually leads to the loss of a group of related attribute data for a continuous period of time, such as data loss caused by icing of anemometers on WTs [9]. The latter is usually manifested as random missing of multiple feature values on some WTs, which is often caused by communication failure or unexpected errors in the data recording process. Based on the knowledge of on-site experts, the frequency of dispersive missing is relatively low [2].

Fig. 1. Two types of missing SCADA data.

Therefore, based on the continuous missing mode, this paper proposes a two-stage multi-view data recovery architecture, and its main contributions are as follows:

(1) We design a two-stage inference framework combining traditional inference methods with deep learning algorithms, providing a new perspective for inferring missing wind power data.
(2) A multi-view graph learning framework is proposed to capture spatial relationships within features and correlations between features. In terms of space, dynamic spatial correlation and static geographic correlation are considered.

2 System Model and Problem Formulation

Assuming there are $N = \{1, 2, \ldots, n\}$ WTs in a large wind farm, with a total monitoring time period of $T = \{1, 2, \ldots, t\}$, and data features are $D = \{1, 2, \ldots, d\}$. The monitoring data of the turbine n is denoted as $X^n_{D,T}$. Thus, the dataset of the wind farm $X^N_{D,T} \in \mathbb{R}^{N \times D \times T}$ during monitoring cycles T is denoted as:

$$X^N_{D,T} = \langle X^1_{D,T}, X^2_{D,T}, \ldots, X^n_{D,T} \rangle = \begin{bmatrix} x^1_{D,1} & \cdots & x^n_{D,1} \\ \vdots & \ddots & \vdots \\ x^1_{D,t} & \cdots & x^n_{D,t} \end{bmatrix} \tag{1}$$

For each node n in the wind farm, we represent its observed value at timestamp t as $x_{d,t}^n \in \mathbb{R}$. To describe the pattern of missing data, we suppose a matrix M that has the same size as X, denoted as $M \in \mathbb{R}^{N \times D \times T}$, where $m_{d,t}^n = 0$ if observation $x_{d,t}^n$ is missing, otherwise $m_{d,t}^n = 1$. Thus, the actual collected missing dataset in the wind farm can be represented as:

$$X^M = X \cdot M \tag{2}$$

where $\cdot$ denotes element-wise product. Then, we use the inference algorithm $\Psi()$ to infer missing data within short periods, with X representing the inferred data. ε represents the error between the observed data and the inferred data.

$$\Psi(X^M) = \widetilde{X} \approx X \tag{3}$$

$$\varepsilon(\widetilde{X}, X) = \sum_{k=1}^{d} \sum_{i=1}^{n} \sum_{j=1}^{t} \left| x_{k,j}^i - \widetilde{x}_{k,j}^i \right| \tag{4}$$

Inference of missing SCADA data is defined as a supervised learning task aimed at minimizing the difference between estimated values and actual values. Assuming the values of missing SCADA data are real numbers, the classification of missing data values is treated as a regression problem.

3 Data Inference Model

3.1 Overall Structure

In a wind farm, the SCADA data from different turbines exhibit high correlation, thus we assume a graph structure to model the spatial correlation among SCADA data. Due to the dynamic nature of SCADA data, this paper utilizes a dynamic weighted graph $G = (V, \varepsilon, \hat{A})$ to accurately describe the interaction within a region, where V is the set of N nodes, each corresponding to a turbine, ε is the set of edges, and $\hat{A} = (A_c, A_d)$ represents the adjacency matrix, consisting of a geographically-based static adjacency matrix A_c and a dynamic adjacency matrix A_d based on mutual information (MI).

The calculation of elements in the static adjacency matrix A_c is as follows:

$$a_{i,j}^c = \begin{cases} e^{-\left(\frac{d_{i,j}}{\delta}\right)^2} & \text{if } a_{i,j}^c < \tau_c \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $d_{i,j}$ represents the Euclidean distance between node i and j, δ is the distance standard deviation, and τ_c is the distance threshold. In the above equation, τ_c controls the threshold for adjacency matrix distribution and sparsity. Distance-based adjacency matrices can capture spatial correlations to some extent because data from nearby nodes typically have similar patterns. However,

considering that data from distant nodes may also exhibit similar trends, distance should not be the sole indicator for determining spatial correlations. The distance-based static adjacency matrix ignores the dynamic spatial correlation between two nodes. MI can be used to estimate the trend of spatial correlation between two nodes. Based on this, the calculation of the element in the dynamic adjacency matrix A_d is as follows:

$$a_{i,j}^{d} = \begin{cases} MI(v_i^{t_d}, v_j^{t_d}) & \text{if } a_{i,j}^{d} < \tau_d \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where v_i and v_j are two nodes in the graph structure, t_d is the period for computing MI. τ_d is the threshold controlling the distribution and sparsity of the dynamic adjacency matrix. A smaller value of $a_{i,j}^{d}$ indicates a smaller difference in feature data between the two nodes during the calculation period.

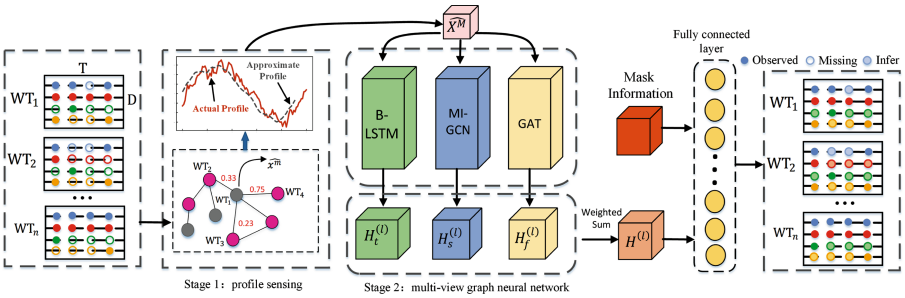

Fig. 2. Overall framework of the proposed methodology

Missing data in deep learning is typically filled with zeros, but these data can somewhat affect feature extraction by neural networks. Therefore, we propose a two-stage method for inferring missing SCADA data. The framework of the proposed method is illustrated in Fig. 2. Firstly, the TS-KNN algorithm is used to fill in missing data at a coarse granularity, giving an approximate profile $\widehat{X^M}$. Then, we use multi-view graph neural networks to extract latent representations aiming at precise prediction of missing data.

3.2 Profile Sensing

We proposed a time series K-nearest neighbors algorithm (TS-KNN) to perform coarse-grained preprocessing of missing data, extracting subsequences $x_{d,[t:t+L]}^{u} = \{x_{d,t}^{u}, x_{d,t+1}^{u}, x_{d,t+2}^{u}, \ldots, x_{d,t+L-1}^{u}\}$ of length L from $x_{d,T}^{u}$. For the subsequences $x_{d,[t:t+L]}^{u}, x_{d,[t:t+L]}^{v}$ in X, their normalized distance is measured by the statistical Pearson correlation coefficient as follows:

$$d_{u,v} = \sqrt{\frac{2L\,|1 - (\Lambda_{u,v} - L\mu_u\mu_v)|}{L\sigma_u\sigma_v}} \tag{7}$$

where $\Lambda_{u,v}$ represents the dot product of corresponding elements in the two subsequences, μ_u and μ_v represent their means, and σ_u and σ_v represent their variances. We calculate the similarity of data for different WTs over the L time points as the distance measure for KNN. We select the k WTs with the smallest distances and take their average as the coarse-grained inferred value for the missing data.

3.3 Multi-view Graph Network Model

After Profile sensing, we propose a multi-view graph network to learn latent representations for specific nodes and time steps. We then use the weighted latent representations of various features to infer missing values of target features. Depending on the considered feature dimensions, it can be divided into three parts: dynamic spatial correlation, inter-feature and spatial correlation at a specific time, and temporal correlation.

Dynamic Spatial Correlation. The traditional spatial Graph Convolutional Network (GCN) structure typically assumes that the receptive field remains unchanged, focusing only on the directly connected first-order neighboring nodes. On one hand, this overlooks those indirectly connected graph nodes and their influence. In fact, not only do first-order nodes exhibit strong spatial proximity, but also those graph nodes relatively far from the target node also show spatial correlations. We proposed a graph convolutional network based on MI dynamic information (MI-GCN). The input of the graph convolutional network is the hidden layer features of the previous graph convolution, along with the weighted adjacency matrix. Utilizing Chebyshev polynomials to approximate the convolution process in GCN improves the traditional receptive field size extension, not limited to directly connected nodes. The minimal unit computation representation of the spatial GCN is as follows:

$$A = U_{c,d}A_c + (1 - U_{c,d})A_d \tag{8}$$

$$T_k(A) = \begin{cases} 2AT_{k-1}(A) - T_{k-2}(A), & \text{if } k \geq 2 \\ A, & \text{if } k = 1 \\ I, & \text{if } k = 0 \end{cases} \tag{9}$$

$$H_s^l = \sigma(\sum_{k=1}^{K} T_k(A)H_s^{(l-1)}\theta^{(l-1,k)}) \tag{10}$$

where A is the weighted adjacency matrix, l is the number of hidden layers, and $U_{c,d}, W, \theta$ are weight matrices formed by neural network parameters.

Through the above design, not only the relationship between dynamic data and static data in spatial structure can be balanced well, but also the adaptability of the model can be enhanced, thereby it is better to capture deep spatial features.

Inter-Feature and Spatial Correlation at a Specific Time. Different attributes of wind power data may demonstrate Different correlations. For example, wind speed and power often exhibit similar trends, while the trend between power and Nacelle temperature may be different. Although the previous view captured spatial correlations between the same attribute data, It did not model the spatial correlations of different attribute data at specific times. Since the dynamic adjacency matrix A_d obtained from the MI layer is learned from all features, it cannot capture feature-specific correlations between nodes. Therefore, we suggest learning the correlations of different attribute data between nodes from the target time series.

Therefore, we design a Graph Attention Network (GAT) based on static adjacency matrix to capture feature-related correlations. The computational expression is as follows:

$$\alpha_{i,j} = \frac{\exp(\text{LeakyReLU}(A_c))}{\sum_{j \in N} \exp(\text{LeakyReLU}(A_c))} \tag{11}$$

$$h_i^{(l)} = \sigma\left(\sum_{j \in N} \alpha_{i,j} W h_j^{(l-1)}\right) \tag{12}$$

$$H_f^{(l)} = \left[h_1^l, h_2^l, \ldots, h_n^l\right] \tag{13}$$

where $H_f^{(l)}$ represents the overall hidden layer features extracted by the graph attention mechanism, and W denotes the learnable spatial and feature weight matrix. In this way, the graph attention mechanism not only captures feature correlations from a longitudinal perspective but also captures spatial correlations at specific times.

Temporal Correlation. Existing spatiotemporal inference tasks usually use variants of RNN to capture temporal dependence. Through empirical comparisons, we choose bidirectional LSTM(B-LSTM), which achieves relatively good performance in our specific Scenario. We apply B-LSTM separately and identically to each node. At each time step, The B-LSTM network recursively takes X_t as the input vector and aggregates historical and future information into two hidden state vectors $Z_{f,t}$ and $Z_{b,t}$, then integrates the learned features from both directions, using

$$Z_{f,t} = LSTM(X_t, Z_{f,t-1}, C_{f,t-1}) \tag{14}$$

$$Z_{b,t} = LSTM(X_t, Z_{b,t-1}, C_{b,t-1}) \tag{15}$$

$$H_t^{(l)} = W_l[Z_{f,t} : Z_{b,t}] + b_l \tag{16}$$

where $Z_{f,t}$, $C_{f,t}$ are the cell memory state vector and candidate state vector in the forward direction, $Z_{b,t}$, $C_{b,t}$ are state vectors in the backward direction, W_l is the parameter matrix for linear transformation, b_l is the biased term.

We combine the latent representations generated by the three modules mentioned above. Because these representations capture correlations from different perspectives and are defined in different hidden spaces, we integrate feature

information from these different dimensions by weighting them and connecting them with some other information, then we infer missing values through a fully connected layer.

$$H^{(l)} = \alpha_1 H_s^{(l)} + \alpha_2 H_t^{(l)} + \alpha_3 H_f^{(l)} \tag{17}$$

$$\widetilde{X} = FC(H^{(l)}||\text{Embedding}(M)) \tag{18}$$

where $\alpha_1, \alpha_2, \alpha_3$ are hyperparameters, $H_t^{(l)}, H_s^{(l)}, H_f^{(l)}$ are features extracted from different dimensions, M is mask matrix, and FC(.) represents a fully connected layer.

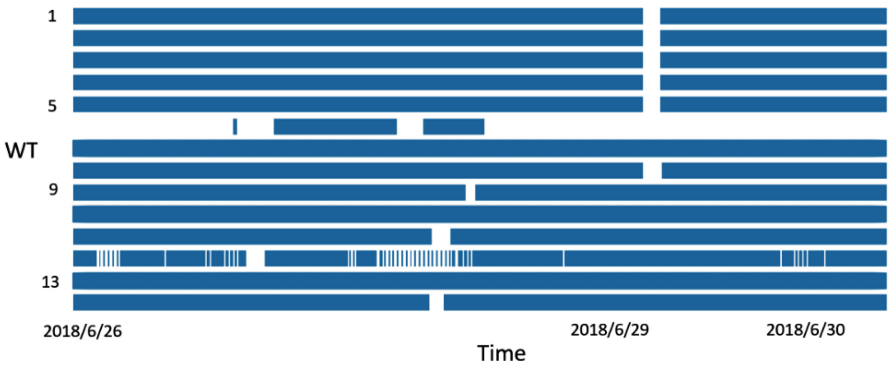

Fig. 3. Missing-data distribution of wind power. Other attributes have similar distributions. The white gaps represent the missing values. The bigger the gap, the more missing values exist.

4 Experiments and Analysis

4.1 Dataset and Experiment Setting

The experimental data is collected from the SCADA system of the Penmanshiel Wind Farm in Ireland, which contains a total of 14 WTs. The data cover the period from January 1, 2017, to July 1, 2021, with a time resolution of 10 min. The dataset comprises three different types of feature data: wind speed, wind power, and Nacelle temperature. There are some real missing value in the dataset such as depicted in Fig. 3.

To validate the effectiveness of the method, The missing data is obtained by randomly deleting some time series of length over 50 from the complete data, making the missing pattern more closely resemble real-world scenarios. We choose h = 6 (i.e., 1 h) for the datasets, where h is the length of time for each sample. We use a sliding-window approach on: $[t, t+h), [t+h, t+2h), [t+2h, t+3h)$, etc. Ensuring that there are no duplicate values in different samples.

4.2 Evaluation Metric and Compared Methods

We use the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) to evaluate our algorithm, which is defined as follows:

$$\text{MAPE} = \frac{1}{dnt}\sum_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{t}\frac{|x_{k,j}^{i} - \hat{x}_{k,j}^{i}|}{x_{k,j}^{i}} \tag{19}$$

$$\text{RMSE} = \sqrt{\frac{1}{dnt}\sum_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{t}(x_{k,j}^{i} - \hat{x}_{k,j}^{i})^2} \tag{20}$$

where $x_{k,j}^{i}$ and $\hat{x}_{k,j}^{i}$) represent the true and inferred values of feature k for turbine i at time point j, respectively. In the experiment, both traditional and state-of-the-art imputation methods are used as our comparison methods.

(1) ST-lazy [4]: ST-lazy algorithm is a classic traditional data inference algorithm. It learns to compute the missing values for the nth WT based on the temporal and spatial correlations between other wind turbines.
(2) IGNNK [13]: IGNNK is a data inference algorithm based on diffusion graph convolution. It is used to recover data from unsampled sensors on graph structures and has shown good performance on many datasets.
(3) GNN-LSTM [14]: GNN-LSTM is a classic data-driven spatiotemporal inference algorithm.

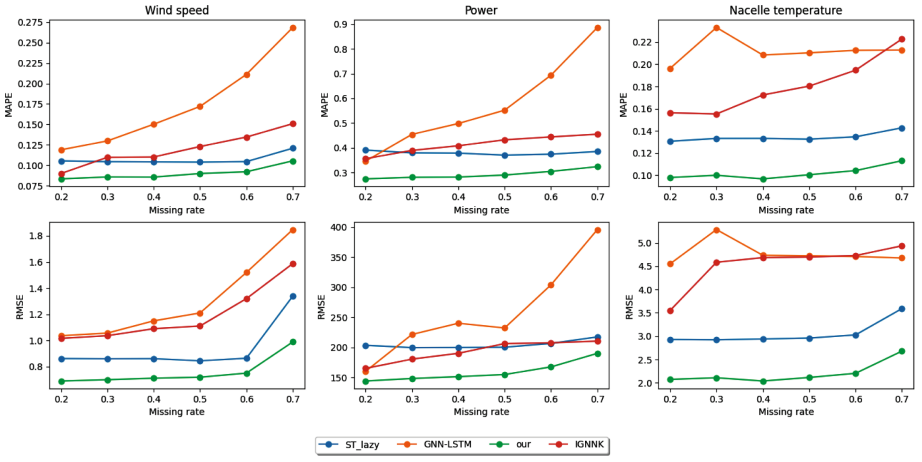

Fig. 4. Imputation error comparison for different methods under different missing rates.

4.3 Imputation Results

Method Performance Under Different Missing Rates. To evaluate the differences in the inference capabilities of our model and the compared method for the task, we use a mask matrix to randomly mask 20%, 30%, 40%, 50%, 60%, and 70% of the values in the SCADA data according to continuous missing patterns. Figure 4 illustrates the performance of our model with three other methods that perform well in the field of missing value imputation. We found that our model outperforms the other contrastive methods across different missing rates. Overall, the ST-lazy algorithm exhibits relatively stable performance across different missing rates and outperforms some deep learning algorithms. This phenomenon may occur because of the relatively strong spatial correlations between nodes, which only impact the ST-lazy algorithm when the missing rate is high. The poor inference performance of the GNN-LSTM algorithm may be attributed to the fact that GNN only performs neighborhood message passing. In wind power fields, spatial relationships are complex and sometimes extend beyond the neighborhood, which GNN fails to capture the changes in global spatial correlations.

The performance of the model may vary across different types of SCADA data. For wind speed and nacelle temperature data, our method still achieved inference errors of 9.21% and 10.42% under a 60% missing rate. across different missing rates, which are significantly better than other algorithms. However, the manual shutdowns of turbines within wind farms, coupled with the occurrence of sudden power fluctuations, generally result in a high inference error of the aforementioned methods for inferring power data.

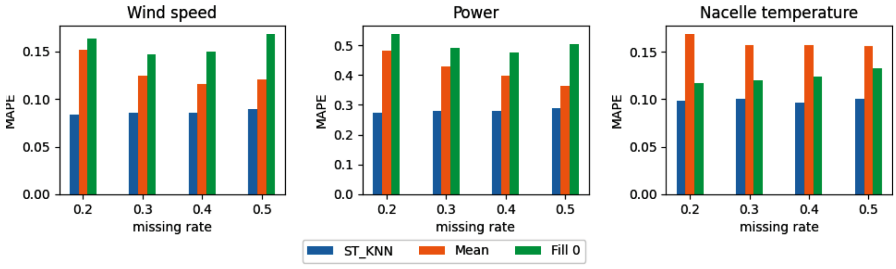

Fig. 5. The impact of different algorithms on the model in profile sense.

Comparison with Different Methods of State1 in Our Method. To better understand the effectiveness of Profile sensing, we adopt different approaches to handle missing values as inputs to the multi-view graph neural network. We observe the performance of the multi-view graph neural network under missing rates of 20%, 30%, 40%, and 50%. As shown in Fig. 5, compared to no filling and mean filling methods, the MAPE estimated by TS-KNN for missing values

significantly decreases. This result validates the motivation behind our proposed TS-KNN Profile sensing method. This approach not only achieves coarse-grained data imputation but also facilitates the calculation of dynamic information for multi-view graph neural networks.

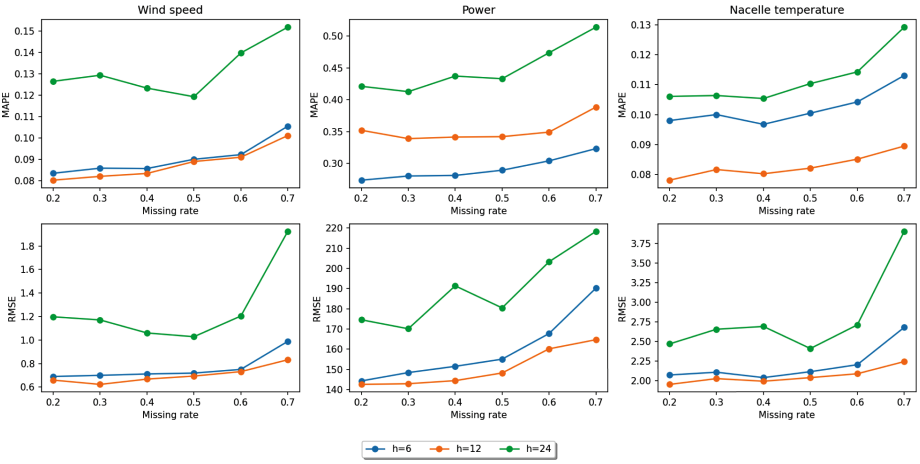

Fig. 6. The impact of different h on method performance.

The Impact of Different Lengths of Time for Sample h on Method Performance. Figure 6 shows the impact of different lengths of time for sample h on our method performance. We used h = 6, 12, and 24, respectively. A large h may lose time dependencies, but a small h might fail to capture the correlations between time steps. In Fig. 6, the inference performance is optimal when h is 12, whereas it significantly deteriorates when h is 24. This suggests that overly large h is adverse to the model from capturing temporal dependencies and bringing noise.

5 Conclusion

The two-stage wind power data inference method proposed in this paper improves the accuracy of inferring missing data from two aspects: 1) a TS-KNN algorithm is used to infer data profile and it is helpful to extract significant features from the multi-view neural network, which improves the effectiveness of data calculation. 2) We proposed a Multi-view Graph Network Model to capture the deep dynamic correlation of time series data from different WTs. Finally, the experimental results show that the method reduces the average inference error RMSE by 19.01% under different missing rates compared with the best comparison algorithm, and the data calculated by the method has high accuracy and stability.

Acknowledgements. This work was supported in part by the Beijing Natural Science Foundation (No. JQ21036), the National Natural Science Foundation of China (No. 62293494, No. 62301078, No. 61821001, No. 62271086), the China Postdoctoral Science Foundation underGrant Number GZB20230086, and the Beijing Key Laboratory of Work SafetyIntelligent Monitoring.

References

1. Global Wind Energy Council. Global Wind Report 2023, Sao Paulo (2023). https://gwec.net/globalwindreport2023
2. Liu, X., Zhang, Z.: A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data. IEEE Sens. J. **21**, 10933–10945 (2021)
3. Li, Z., et al.: A spatiotemporal directed graph convolution network for ultra-short-term wind power prediction. IEEE Trans. Sustain. Energy **14**, 39–54 (2022)
4. Sun, C., Chen, Y., Cheng, C.: Imputation of missing data from offshore wind farms using spatio-temporal correlation and feature correlation. Energy **229**, 120777 (2021)
5. Poloczek, J., Treiber, N., Kramer, O.: KNN regression as geo-imputation method for spatio-temporal wind data. In: Proceedings of the International Joint Conference SOCO 2014-CISIS 2014-ICEUTE 2014, Bilbao, 25–27 June 2014, pp. 185–193 (2014)
6. Fouladgar, N., Främling, K.: A novel LSTM for multivariate time series with massive missingness. Sensors **20**, 2832 (2020)
7. Marisca, I., Cini, A., Alippi, C.: Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. Adv. Neural. Inf. Process. Syst. **35**, 32069–32082 (2022)
8. Kuppannagari, S., Fu, Y., Chueng, C., Prasanna, V.: Spatio-temporal missing data imputation for smart power grids. In: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, pp. 458–465 (2021)
9. Coville, A., Siddiqui, A., Vogstad, K.: The effect of missing data on wind resource estimation. Energy **36**, 4505–4517 (2011)
10. Li, H., Liu, L., He, Q.: A joint missing power data recovery method based on the spatiotemporal correlation of multiple wind farms. J. Renew. Sustain. Energy **16** (2024)
11. Fan, H., Zhang, X., Mei, S.: Wind power time series missing data imputation based on generative adversarial network. In: 2021 IEEE 4th International Electrical and Energy Conference (CIEEC), pp. 1–6 (2021)
12. Hu, X., Zhan, Z., Ma, D., Zhang, S.: Spatiotemporal generative adversarial imputation networks: an approach to address missing data for wind turbines. IEEE Trans. Instrument. Measur. (2023)
13. Wu, Y., Zhuang, D., Labbe, A., Sun, L.: Inductive graph neural networks for spatiotemporal kriging. Proc. AAAI Conf. Artif. Intell. **35**, 4478–4485 (2021)
14. Li, J., et al.: A nested machine learning approach to short-term PM2. 5 prediction in metropolitan areas using PM2. 5 data from different sensor networks. Sci. Total Environ. **873**, 162336 (2023)