Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 4170064, 15 pages
https://doi.org/10.1155/2021/4170064

Research Article

Joint Matrix Decomposition-Based Missing Data Completion in Low-Voltage Area

Haowen Wu,[1] Chen Yang,[1] Wenwang Xie,[1] and Wei Zhang[2]

[1]*China Southern Power Grid Digital Grid Research Institute Co., Ltd., Guangzhou, Guangdong 510663, China*
[2]*Hefei University of Technology, Hefei, Anhui 230009, China*

Correspondence should be addressed to Wei Zhang; 2020170417@mail.hfut.edu.cn

Received 23 August 2021; Revised 30 September 2021; Accepted 18 October 2021; Published 8 November 2021

Academic Editor: Xianyong Li

In-depth mining and analysis of electricity data in low-voltage area are essential for the further intelligent development of power grids. However, in the actual data collection and measurement of low-voltage area, there will be missing data, and complete electricity data cannot be obtained. To obtain complete power data, this paper proposes a low-voltage station area missing data complement model based on joint matrix decomposition. First, we analyse the characteristics of the low-pressure station data. Then, a model that comprehensively considers the characteristics of the low-voltage station area data is proposed, which includes three parts: the construction of a low-voltage station area data tensor, the joint matrix decomposition, and the completion of the missing data, and it is named LPZ. After that, the CIM learning algorithm proposed in this paper is used to iteratively solve the model to obtain the completed data. Finally, the method proposed in this paper is used to complement the two situations of random loss and all-day loss of real current data in a low-voltage station area and compared with the traditional complement method. The experimental results show that this method is not only effective but also that the completion effect is better than that of other completion methods.

1. Introduction

In recent years, with the continuous advancement of intelligent power grid construction, in-depth mining and analysis of electricity data have become increasingly important [1, 2]. Electricity data contain a large amount of electricity consumption data information. Through in-depth mining and analysis of electricity consumption data, various advanced applications, such as electricity demand and electricity price setting, can be realized to provide support for the safe and efficient operation of the power grid [3–5]. As an important part of the power grid, the low-voltage station area, in-depth mining, and analysis of its electricity data will become the key to further intelligentization of the power grid, which has important significance for the future [6].

To successfully realize the in-depth mining and analysis of the electricity data of the low-voltage station area, it is necessary to maintain the integrity of its electricity consumption data as much as possible. In the actual process of the low-voltage station area, the data are missing due to equipment damage, weather conditions and other reason, causing a sharp fall in the quality. For example, the data in the literature [7] have a missing rate of multiple attributes exceeding 50%, and there are missing data almost in every record. Therefore, in the in-depth mining and analysis of electricity data, it is necessary to process the missing data with the known data [8, 9].

At present, scholars at home and abroad have carried out some researches on the methods of completing missing data. Literature [10] attempted to complement the synchrophasor measurement data of the synchrophasor measuring device using the matrix filling method. However, as low-voltage station area power data have different characteristics from synchronized phasor measurement data, that is, the similarity of the power consumption data of each user is quite different, it is difficult to directly apply the matrix filling theory to achieve the repair effect. Literature [11] utilized an

adaptive neuro-fuzzy inference system model to complement and optimize the missing power data based on the traditional single data completion method (such as interpolation). The aforementioned methods require a large amount of data for pretraining. Thus, a small sample size may lead to unsatisfactory effect of completion. Literature [12] uses the KNN completion algorithm to complete missing values. Although the KNN completion algorithm is simple, intuitive, and easy to implement without the need for prior knowledge, the accuracy of its completion depends on the average value of the neighboring sample data. Literature [13] based on the single-value missing data completion method of nonlocal averaging method proposed a multivalue missing data completion method using spatial neighbor BP (backpropagation) mapping to achieve higher precision data completion. However, this method does not consider the timing of the data to be completed. Literature [14] considered that in the actual system, a subset of the data may have stronger relevance, and proposed a new local tensor completion model. This model uses the stronger local correlation of the data to form and restore each subtensor with a lower level to achieve accurate data recovery. However, this method cannot achieve data completion when there is less correlation between the data. Literature [15] adopted the machine learning-based method to perform missing value completion. Although this type of method performs well in accuracy, it is necessary to learn the complete data sequence, and it is difficult to find the complete sequence training parameters in actual search.

Given the above problems, this paper proposes a low-voltage station data completion method based on joint matrix decomposition based on the characteristics of the low-voltage station area electricity data. First, we analyse the characteristics of the low-pressure station data. Then, a model that comprehensively considers the characteristics of the low-voltage station area data is proposed. The modified model includes three parts: the construction of a low-voltage station area data tensor, joint matrix decomposition, and completion of the missing data, and it is named LPZ. After that, the CIM learning algorithm proposed in this paper is used to iteratively solve the model to obtain the completed data. Finally, through the verification of real current data of a low-voltage station area, the effectiveness of the method proposed in this paper is obtained, and compared with the traditional complement method, the superiority of the method in this paper is obtained.

2. Analysis of the Characteristics of Electricity Data in Low-Voltage Area

The low-voltage station area electricity data mainly include the voltage, current, active power, reactive power, and other data of each user in the station area [16]. The data mainly have the following characteristics:

(1) *Periodicity*. Generally, the power data of low-voltage area shows periodic changes over several consecutive working days; that is, on consecutive working days, the electricity consumption behavior of each user shows a similar periodic law. As shown in Figure 1, on different working days, the current curve of a user in the station area has a similar trend.

(2) *Sequentiality*. User data all appear in the form of data streams, which are sequentially collected, transmitted, and stored at equal time intervals. The data analysed in this paper are based on a sampling interval of 30 minutes, and 48 points of data are collected a day. The data graph with the sampling interval as the time window is shown in Figure 2.

(3) *Spatial Correlation*. In the power system, different users are connected through the network topology of the station area, and the power load between different users has a certain correlation, especially when a high-power electrical appliance starts or malfunctions. The performance will be more obvious. Therefore, it is necessary to consider the multiuser spatial correlation of the station area data to complete the missing data of multiple users.

3. Complementary Method for Missing Data in Low-Voltage Station Areas Based on Joint Matrix Decomposition

3.1. Model Structure. According to the analysis of the characteristics of low-voltage station area electricity data, we conclude that the low-voltage station area data contain three characteristics: periodicity, time series, and spatial correlation. Therefore, we propose a model that comprehensively considers the data characteristics of the low-voltage station area to solve the problem of data completion in the low-voltage station area and name it LPZ. Specifically, first, we design the organization of electricity data. The electricity consumption data sequence of all users in the low-voltage station area is organized into a tensor, which can not only mine potentially related information from different patterns but also ensure the original characteristics of the station area electricity data. Next, we propose a joint decomposition module [17] that extracts the day-time interval matrix and the user-time interval matrix from this tensor and then decomposes all the extracted matrices, that is, an original matrix is expressed in the form of the product of two low-dimensional matrices to form an expression of all users, days, and time intervals. Furthermore, to mine the characteristics of the temporality of the data, we added a local restriction to the joint decomposition module. The restriction condition we adopted here is to make the predicted value of the target value close to the predicted value of the adjacent time interval.

The LPZ architecture is shown in Figure 3, including three parts: construction of a low-voltage station area data tensor, joint matrix decomposition, and missing data completion. Among them, the two decomposition modules in the joint decomposition matrix provide the characteristic expression of users, days, and time intervals by mining potential factors, and both decomposition modules are affected by local restrictions, which makes the model consider the period of low-voltage station data. It is also possible to

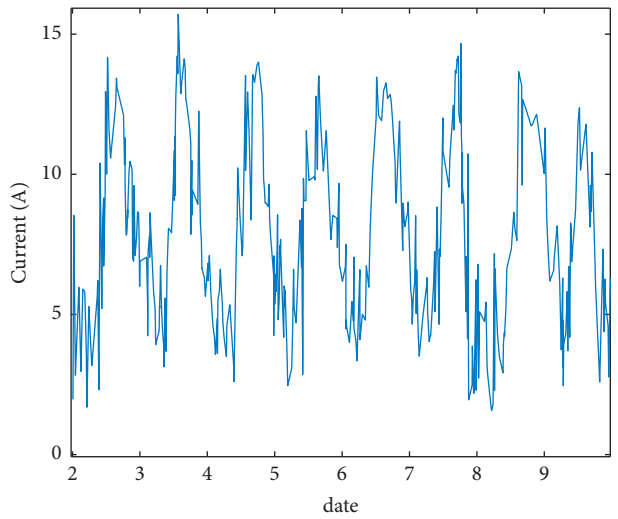

Figure 1: Current curve of a user for 8 consecutive days.

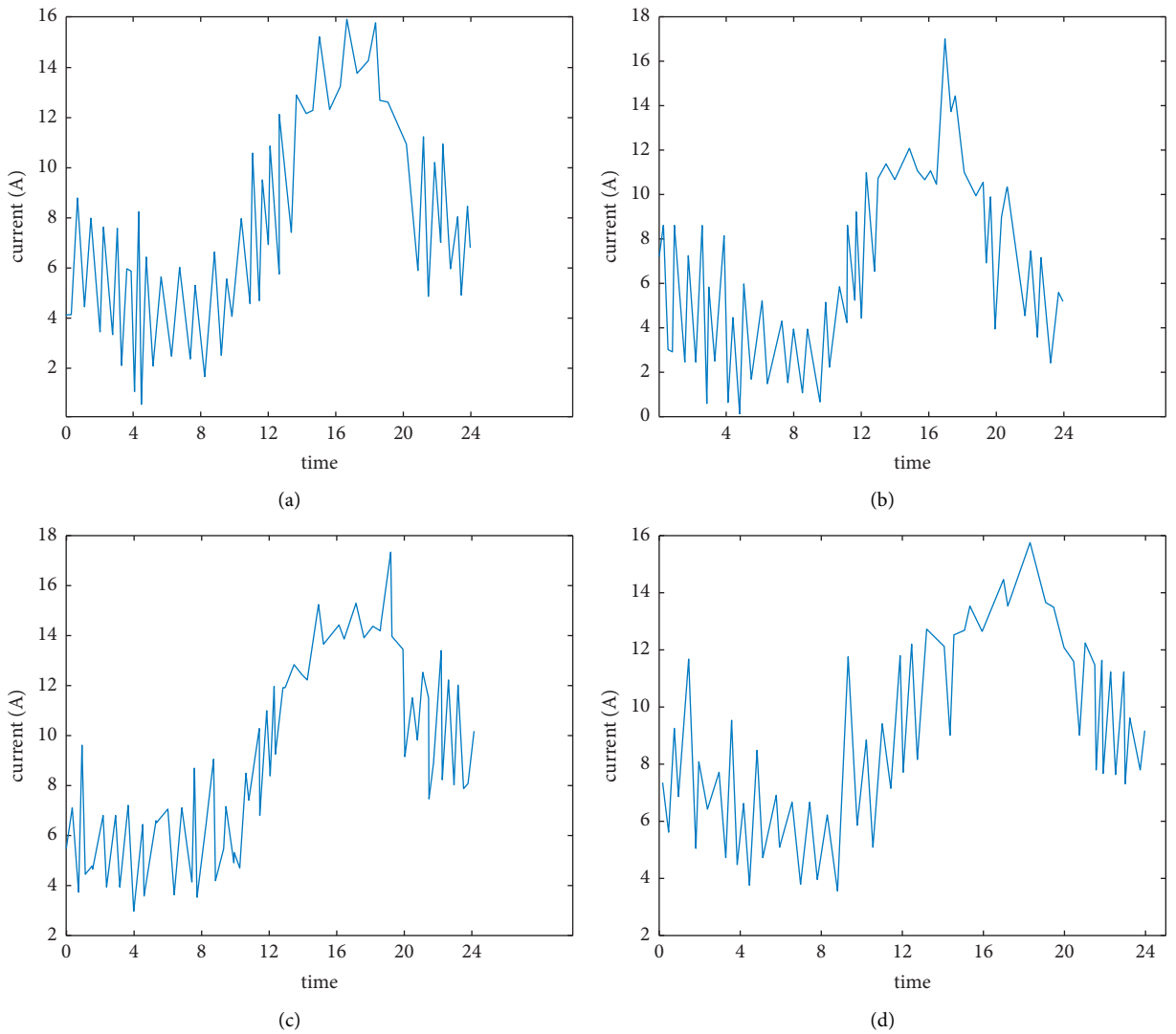

Figure 2: The current size of the same user on four days in a month. (a) First day's data. (b) Fourth day's data. (c) Eighth day's data. (d) Twelfth day's data.

mine and utilize the temporality of data as well as spatial and spatial relevance.

3.2. Constructing the Data Tensor of the Low-Pressure Station Area. Based on the characteristics of low-voltage station area electricity data, we designed the input of the model. First, we assume that the power consumption data of user i in time interval j on day k is $c_{i,j,k}$; then, the power consumption data sequence obtained by user i in chronological order is $S_i = <c_{i,0,0}, \ldots, c_{i,j,0}, c_{i,0,1}, \ldots, c_{i,J,1}, \ldots, c_{i,J,K}>$, where J is the number of time intervals in a day and K represents the number of days. We fold it into K vectors, and each vector contains the electricity consumption data of the user at various time intervals in a day. Then, these vectors are integrated to form a two-dimensional matrix form, as shown in equation:

$$\begin{bmatrix} c_{i,0,0} & c_{i,1,0} & \cdots & c_{i,J,0} \\ c_{i,0,1} & c_{i,1,1} & \cdots & c_{i,j,1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{i,0,K} & c_{i,1,K} & \cdots & c_{i,J,K} \end{bmatrix} \in R^{K\times J}. \tag{1}$$

Performing a folding operation on the electricity consumption data sequence of I users can obtain I two-dimensional matrices. Then, we integrate them into a three-dimensional tensor (user, time interval, and day) and use it as input to the model. This input not only maintains the characteristics of the original data but also makes the tensor pattern closely related to the characteristics of the low-voltage station area electricity data.

3.3. Joint Matrix Factorization. Due to the regularity of people's activities, on a continuous working day, a user's electricity consumption data will generally show periodic changes. To model the periodicity of user electricity data, we extract the number of day-time interval matrix $\mathbf{C}_i$ from the electricity data tensor C and learn the temporality from $\mathbf{C}_i$ through matrix decomposition. Then, hidden factors are introduced, and matrix $\mathbf{C}_i$ is decomposed into two low-dimensional matrices $\mathbf{P}_i \in R^{K\times F}$ and $Q_i \in R^{J\times F}$, where F is the number of hidden factors. In addition, each day k is related to a vector $p_k \in R^F$. Similarly, each time interval j corresponds to a vector $q_j \in R^F$, where the similarity between the electricity consumption data of different days or different time intervals can be captured in a potentially low-dimensional space. Thus, the predicted value $c_{i,j,k}$ of the power consumption data $\widetilde{c}_{i,j,k}$ of user i in time interval j of day k is obtained, which is represented by the inner product $p_k \cdot q_j$.

At the same time, the electricity consumption data series of different users will also be correlated. Therefore, we also decompose the user-time interval matrix $\mathbf{C}_k$ to explore the spatial correlation of electricity consumption data. All users and time intervals are mapped to a low-dimensional space. In this latent space, $\mathbf{S}_k \in R^{I\times F}$ represents all users, and $\mathbf{T}_k \in R^{J\times F}$ represents all time intervals. Moreover, each user i corresponds to a vector $s_i \in R^F$, and each time interval j corresponds to a vector $t_j \in R^F$. Thus, the predicted value $\widetilde{c}_{i,j,k}$ of the power consumption data $c_{i,j,k}$ of user i in time interval j of day k is obtained, which is expressed by the inner product $s_i \cdot t_j$. To obtain more accurate prediction results, when predicting the missing low-voltage station data, we combine the two matrix decomposition modules to take periodicity and spatial correlation into account at the same time.

In addition, the electricity consumption data of a certain time interval have a strong correlation with the electricity consumption data of the surrounding time interval. Therefore, we introduce local constraints into the joint matrix factorization process. In the process of decomposing the number of day-time intervals, we also need to minimize the difference between the predicted value of the target electricity consumption data and the mean value of the surrounding time interval, as shown in the following equation:

$$\begin{aligned} g_1 &= \left(p_k \cdot q_j - \overline{c}^{(1)}_{i,j,k}\right)^2, \\ \overline{c}^{(1)}_{i,j,k} &= \frac{1}{2W}\sum_{\omega}^{W}\left(p_k \cdot q_{j-\omega} + p_k \cdot q_{j-\omega}\right), \end{aligned} \tag{2}$$

where W is the window size, and $\omega = 1, 2, \ldots, W$.

Similarly, a local restriction is also added in the decomposition process of the user-time interval matrix, as shown in following equation:

$$\begin{aligned} g_2 &= \left(s_i \cdot t_j - \overline{c}^{(2)}_{i,j,k}\right)^2, \\ \overrightarrow{c}^{(2)}_{i,j,k} &= \frac{1}{2W}\sum_{\omega=1}^{W}\left(s_i \cdot t_{j-\omega} + s_i \cdot t_{j+\omega}\right), \end{aligned} \tag{3}$$

where W is the window size, and $\omega = 1, 2, \ldots, W$.

In summary, the low-voltage station area data contain three characteristics, namely, periodicity, time series, and spatial correlation. In addition, the collected data contain a large number of missing values, causing data sparsity problems. Therefore, we design and use the joint matrix decomposition module to model the periodicity and spatial correlation, respectively. At the same time, we set local constraints based on the spatial correlation to restrict the joint decomposition module. In this way, the three features can work synergistically when completing missing values.

3.4. Completion of Missing Data. Through joint matrix decomposition, we can obtain the hidden factor matrices $\mathbf{P}_i$ and $\mathbf{Q}_i$ of user i and the hidden factor matrices $\mathbf{S}_k$ and $\mathbf{T}_k$ of the number of days k. Finally, we obtain four-parameter tensors, namely, $P \in R^{K\times F\times I}$, $\mathbf{Q} \in R^{J\times F\times I}$, $\mathbf{S} \in R^{I\times F\times K}$, and $\mathbf{T} \in R^{J\times F\times K}$. Using these factor tensors can realize the completion of the original incomplete matrix. Here, we use a simple regression method to combine the two partial results and use it as the output of the joint matrix factorization module. The weight β is set to control this combination process, where the weight β represents the periodic force, and $(1-\beta)$ represents the weight of the spatial correlation in

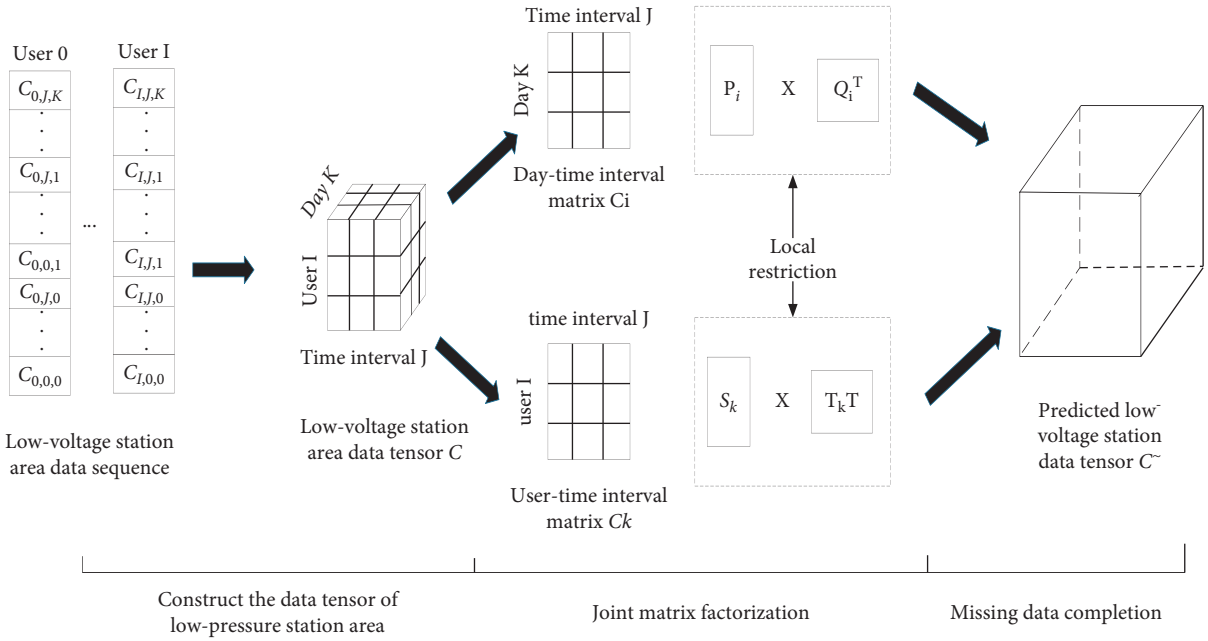

Figure 3: LPZ architecture diagram.

the missing value prediction process. This linear fitting process enables LPZ to reasonably weigh the characteristics of the low-voltage station area data, thereby obtaining a complete low-voltage station area electricity data tensor. Thus, the prediction formula of the missing value shown in formula (4) is obtained:

$$\widetilde{c}_{i,j,k} = \beta P_{k,:,i} Q^T_{j,:,i} + (1-\beta) S_{i,:,k} T^T_{j,:,k}, \tag{4}$$

where $\mathbf{P}_{k,:,i} \in \mathbf{R}^F$ is equivalent to p_k; $\mathbf{Q}_{j,:,i} \in R^F$ is equivalent to q_j; $\mathbf{S}_{i,:,k} \in R^F$ and s_i are equivalent; and $\mathbf{T}_{j,:,k} \in R^F$ is equivalent to t_j.

3.5. Objective Function and Parameter Learning. We establish the model objective function by minimizing the square difference between the true value of the low-voltage station area electricity data and its estimated value. Given a low-voltage station area electricity data tensor C, this tensor contains a large number of missing values. Our model can mine multiple features of the low-voltage station area electricity data and generate one and the original tensor based on the known values in the tensor, completely estimating tensors with the same shape to realize the task of complementing low-pressure station area data.

For simplicity, we set a binary mask $\mathbf{M}$; this mask tensor corresponds to the original tensor, and its value is also determined by the value of the element at the corresponding position in the original tensor. In the masking tensor, the missing element in the original low-voltage station area electricity data tensor has a value of 1, and the observed element position is 0; that is, through the masking tensor $\mathbf{M}$, we can clearly know the missing values in the original tensor, and the position of the observation value is formulated as the following equation:

$$m_{i,j,k} = \begin{cases} 1, & c_{i,j,k} \text{ missing}, \\ 0, & c_{i,j,k} \text{ known}. \end{cases} \tag{5}$$

Therefore, the missing value in the original matrix can be expressed as $\mathbf{M} \odot \mathbf{C}$, and the observed value can be expressed as $(1 - \mathbf{M}) \odot \mathbf{C}$. The objective function can be defined as the following formula:

$$\ell = \|(1-\mathbf{M}) \odot (\widetilde{C} - \mathbf{C})\|_F^2 + \lambda \|\theta\|^2, \tag{6}$$

where $\odot$ represents the point multiplication operation, $\|\cdot\|_F$ represents the Frobenius norm, $\lambda\|\theta\|^2$ represents the regularization term to prevent overfitting, and θ represents all parameter tensors.

According to the description in the previous two subsections, to consider the periodicity and spatial correlation of the low-voltage station area data at the same time, we designed a joint decomposition module and added local restrictive constraints in the decomposition process so that LPZ can consider the electricity data at the same time. Here, we use the linear fitting method to combine the local results of the two decomposition models to obtain the objective function shown in the following equations:

$$
\begin{aligned}
l = & \underbrace{\beta \sum_{i=0}^{I}\sum_{j=0}^{J}\sum_{k=0}^{K}\left(\mathbf{P}_{k,:,i}Q_{j,:,i}^{T}-c_{i,j,k}\right)^{2}}_{Spatial\ correlation} \\
& + \underbrace{(1-\beta)\sum_{i=0}^{I}\sum_{j=0}^{J}\sum_{k=0}^{K}\left(S_{i,:k}T_{j,:,k}^{T}-c_{i,j,k}\right)^{2}}_{Periodic} \\
& + \underbrace{\lambda_1 \sum_{i=0}^{I}\sum_{j=0}^{J}\sum_{k=0}^{K}\left(\mathbf{P}_{k,i}\mathbf{Q}_{j,:,i}^{T}-\overline{c}_{i,j,k}^{(1)}\right)^{2}_{Sequentiality}+\lambda_1 \sum_{i=0}^{I}\sum_{j=0}^{J}\sum_{k=0}^{K}\left(\mathbf{S}_{i,:,k}\mathbf{T}_{j,,,k}^{T}-\overline{c}_{i,j,k}^{(2)}\right)^{2}_{Sequentiality}+\lambda\|\theta\|^{2}}_{},
\end{aligned}
\tag{7}
$$

$$
S.t: \quad \mathbf{P}\geq 0,\ \mathbf{Q}\geq 0,\ \mathbf{T}\geq 0 \text{ and } \mathbf{S}\geq 0. \tag{8}
$$

It should be noted that in the completion process, due to the role of the masking tensor M, only observations are used to train the model. However, to simplify the objective function, in formulas (7) and (8), we omit the masking tensor. According to formula (1), we can find that in the original tensor, a known low-voltage station area electricity data element will affect the four-parameter tensors. In addition, the predicted value is not only the combination of the results of the two decomposition modules but also the constraint of local limitation; that is, the predicted value of the low-pressure station area data is affected by the real value related to it and the predicted value of the surrounding time interval at the same time. The matrix extracted from the original tensor has the problem of data sparsity, which can be solved by using nonnegative matrix factorization, and its nonnegativity ensures the interpretability of the learned parameter tensor.

After clarifying the objective function of LPZ, we will next introduce in detail how to estimate the parameters. Compared with the stochastic gradient descent method [18], the alternating least squares method is easier to adjust and highly parallelizable [19]. Therefore, in this article, we refer to the alternating least squares method for parameter estimation. During the training process, the observations will be used to update the model parameters iteratively until the objective function converges. In one iteration, all samples in the original tensor will be traversed once, and at the same time, the four-parameter tensors will be updated with one sample each time. Here, a known element in the tensor is a training sample.

The parameters to be trained are the four-parameter tensors of the model. Here, we separately calculate the objective function for each vector $\{\mathbf{P}_{k,:,i}, \mathbf{Q}_{j,:,i}, \mathbf{S}_{i,:,k}, \mathbf{T}_{j,:,k}\}$. The partial derivative and the result are shown in the following formulas:

$$
\begin{aligned}
\frac{\partial \ell}{\partial P_{k,:,i}} = & 2\beta \sum_{j=1}^{J}\left(\mathbf{P}_{k,:,i}\mathbf{Q}_{j,:,i}^{T}-c_{i,j,k}\right)-Q_{j,:,i} \\
& + 2\lambda_1 \sum_{j=1}^{J}\left(\mathbf{P}_{k,:,i}\mathbf{Q}_{j,:,i}^{T}-\overline{c}_{i,j,k}^{(1)}\right)\mathbf{D}_{j,:,i}^{(1)}+2\lambda\mathbf{P}_{k,:,i},
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \mathbf{Q}_{j,:,i}} = & 2\beta \sum_{k=1}^{K}\left(\mathbf{P}_{k,:,i}\mathbf{Q}_{j,:,i}^{T}-c_{i,j,k}\right)\left(-\mathbf{P}_{\mathrm{k},:,i}\right) \\
& + 2\lambda_1 \sum_{k=1}^{K}\left(\mathbf{P}_{k,:,i}\mathbf{Q}_{j,:,i}^{T}-\overline{c}_{i,j,k}^{(1)}\right)\mathbf{P}_{k,:,i}+2\lambda\mathbf{Q}_{j,:,i},
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial S_{i,:,k}} = & 2(1-\beta)\sum_{j=1}^{J}\left(\mathbf{S}_{i,:,k}\mathbf{T}_{\mathrm{j},:,k}^{T}-c_{i,j,k}\right)\left(-\mathbf{T}_{j,:,k}\right) \\
& + 2\lambda_1 \sum_{j=1}^{J}\left(\mathbf{S}_{i,:,k}\mathbf{T}_{j,:,k}^{T}-\overline{c}_{i,j,k}^{(2)}\right)\mathbf{D}_{j,:,i}^{(2)}+2\lambda S_{i,:,k},
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial T_{j,:,k}} = & 2(1-\beta)\sum_{i=1}^{I}\left(\mathbf{S}_{i,:,k}\mathbf{T}_{\mathrm{j},:,k}^{T}-c_{i,j,k}\right)\left(-\mathbf{S}_{i,:,k}\right) \\
& + 2\lambda_1 \sum_{i=1}^{I}\left(\mathbf{S}_{i,:,k}\mathbf{T}_{\mathrm{j},:,k}^{T}-c_{i,j,k}^{T}\right)\mathbf{S}_{i,:,k} \\
& + 2\lambda T_{j,:,k}.
\end{aligned}
\tag{12}
$$

Among them, $\mathbf{D}_{j,:,i}^{(1)}$ and $\mathbf{D}_{j,:,i}^{(2)}$ are auxiliary variables. The tensor $\mathbf{D}^{(1)} \in R^{J\times F\times I}$ corresponds to $\mathbf{Q}$. $\mathbf{D}_{j,:,i}^{(1)} \in R^{F}$ is the difference between the expression of time interval j and the mean value of the surrounding time interval. Its formula is shown in (13). Similar to $\mathbf{D}^{(1)}$, the tensor $\mathbf{D}^{(2)}$ is also used to describe the difference in time interval expression, and $\mathbf{D}_{j,:,k}^{(2)} \in \mathbf{R}^{F}$. The difference is that $\mathbf{D}^{(2)} \in \mathbf{R}^{J\times F\times I}$ corresponds to $\mathbf{T}$, and the time interval expression used by $\mathbf{D}_{j,:,k}^{(2)}$ is

generated based on spatial correlation. The formulaic representation is shown in (14).

$$\mathbf{D}_{j,:,i}^{(1)} = \mathbf{Q}_{j,:,i} - \frac{1}{2W}\sum_{\omega=1}^{W}\left(\mathbf{Q}_{j-\omega,:,i} + \mathbf{Q}_{j+\omega,:,i}\right), \tag{13}$$

$$\mathbf{D}_{j,:,k}^{(2)} = \mathbf{T}_{j,:,k} - \frac{1}{2W}\sum_{\omega=1}^{W}\left(\mathbf{T}_{j-\omega,:,k} + \mathbf{T}_{j+\omega,:,k}\right), \tag{14}$$

where W is the window size, and $\omega = 1, 2, \ldots, W$, that is, the power data of a certain time interval will be affected by the time interval before and after it.

The process of parameter update is iterative. Referring to the update process of the alternating least squares matrix decomposition, we set the gradient to 0 to derive the parameter update formula, as shown in the following equations:

$$\mathbf{P}_{k,:,i} = \left[\beta\mathbf{Q}_i^T\mathbf{Q}_i + \lambda_1\mathbf{D}_i^{(1)T}\mathbf{D}_i^{(1)} + \lambda\mathbf{I}\right]^{-1} \cdot \left[\beta\sum_{j=1}^{J} c_{i,j,k}\mathbf{Q}_{j,:,i}\right], \tag{15}$$

$$\mathbf{Q}_{j,:,i} = \left[\beta\mathbf{P}_i^T\mathbf{P}_i + \lambda_1\mathbf{P}_i^T\mathbf{P}_i + \lambda\mathbf{I}\right]^{-1} \cdot \left[\beta\sum_{k=1}^{K} c_{i,j,k}\mathbf{P}_{k_i,i} + \lambda_1\sum_{k=1}^{K}\overline{c}_{i,j,k}^{(1)}\mathbf{P}_{k,:,i}\right], \tag{16}$$

$$S_{i,:,k} = \left[(1-\beta)\mathbf{T}_k^T\mathbf{T}_k + \lambda_1\mathbf{D}_k^{(2)T}\mathbf{D}_k^{(2)} + \lambda\mathbf{I}\right]^{-1} \cdot \left[(1-\beta)\sum_{j=1}^{J} c_{i,j,k}\mathbf{T}_{j,:,k}\right], \tag{17}$$

$$\mathbf{T}_{j,:,k} = \left[(1-\beta)\mathbf{S}_k^T\mathbf{S}_k + \lambda_1\mathbf{S}_k^T\mathbf{S}_k + \lambda\mathbf{I}\right]^{-1} \cdot \left[(1-\beta)\sum_{j=1}^{J} c_{i,j,k}\mathbf{S}_{i,:,k} + \lambda_1\sum_{i=1}^{I}\overline{c}_{i,j,k}^{(2)}\mathbf{S}_{i,:,k}\right]. \tag{18}$$

Specifically, in an iterative update, the tensors **Q**, **S**, and **T** are fixed first, and the row vector of the tensor **P** is updated according to the above formula. After updating of the **P** is completed, **P**, **S**, and **T** are fixed to update **Q** row by row. Next, **S** and **T** are updated separately in this update mode. Obviously, because for an update of a parameter tensor, the update between the row vectors does not affect each other, this process is highly parallelizable, which can greatly speed up the training speed of the model. In one iteration, the four-parameter tensors will be updated separately according to formulas (15)–(18) and continue until the objective function converges. It is worth noting that this update method does not guarantee the nonnegativity of the parameter tensors **P**, **Q**, **S**, and **T**. Because our objective function is continuous, its minimum value should be obtained at the point where the gradient is 0 or the point on the boundary. In this paper, a simple method is used to deal with negative values in the parameter tensor. If there is a value less than 0 in the parameter tensor, set it to 0.

Figure 4 summarizes the training process of the LPZ model. First, the original low-voltage station area electricity data tensor is formed and used as the input of the model. Then, the model is trained and the four-parameter tensors are iteratively updated until the objective function converges (lines 2–8). Finally, the results of the decomposition modules are averaged in a weighted manner and used as the predicted values of the missing values (lines 9–13).

4. Experiment and Analysis

For the validity of our model, we conducted a large number of experiments on a dataset of a certain station in a distribution network of a certain city in China. The LPZ is compared with three current data complementation methods, and the experimental results show that LPZ can obtain better prediction results than the current complementation methods. In this section, we first introduce the dataset and experimental settings; second, we use different parameters to evaluate the LPZ model; and finally, we conduct comparative experiments and analyse the experimental results.

4.1. Dataset and Experimental Settings. This section first introduces the dataset used in the experiment and then explains the experimental parameter settings and evaluation indicators.

4.1.1. Dataset. For the dataset of this experiment, we will use the user current data of a certain station in the distribution network of a certain city in China. The station structure is shown in Figure 5. Here, VLV22 represents the cable model, and 4×70 represents 4-core 70 mm^2. In the actual data collected automatically, the current data of a certain month are randomly selected as the data test set to construct the current tensor. The constructed current tensor contains a total of 142650 ($317\times180\times25$) elements, of which the number of nonzero elements is 1209011. We randomly selected 80% of the nonzero elements as the training set and the remaining 20% as the test set to prove the effectiveness of our proposed model.

4.1.2. Parameter Setting. In this section, the parameter settings of the LPZ model are mainly discussed, and these default settings are obtained through parameter tuning. In the experiment, we set the default value of the number of hidden factors **F** to 15; that is, we use a 15-dimensional vector to represent users, time intervals, and days. The weight parameter β is used to control the combining process of the partial results of the two subdecomposition modules, and its default setting is 0.4. The default window size W is 4; that is, the current data of a time interval are affected by 4-time intervals before and after it. In the local limit, λ_1 is used to control the influence of sequentiality, and the default value is 0.1. At the same time, for the regularization term coefficient λ, the default value is also set to 0.1.

Algorithm 1: CIM learning algorithm

Input: The original low-pressure station data tensor with missing values $\boldsymbol{C}$, mask tensor $\boldsymbol{M}$, number of hidden factors $\boldsymbol{F}$ and the parameter tensors $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{S}$ *and* $\boldsymbol{T}$;

Output: Fully estimated tensor $\widetilde{\boldsymbol{C}}$;

Initialize $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{S}$ *and* $\boldsymbol{T}$; // Training model

repeat

for $c_{i,i,k} \in \boldsymbol{C}$ **do**

if $m_{i,j,k} = 0$ **then**

Update the parameter tensor according to formula (15)-(18);

end if

end for

until the stop condition // prediction is met

for $m_{i,j,k} \in M$ **do**

if $m_{i,j,k} = 1$ **then**

Calculate missing values according to formula (4)

end if

end for

return fully estimated tensor $\widetilde{\boldsymbol{C}}$;

Figure 4: Training algorithm of the LPZ model.

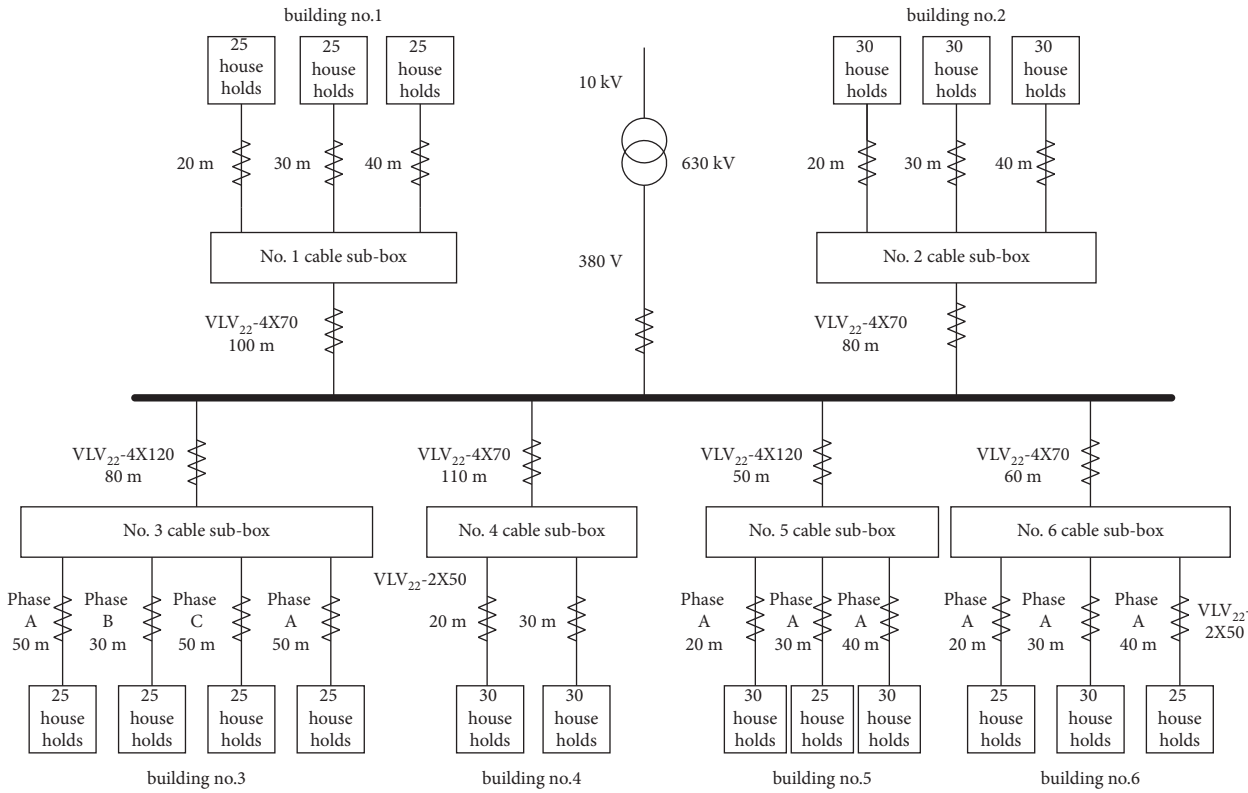

Figure 5: Structure diagram of the low-voltage station area.

4.1.3. Evaluation Index. In this article, we use root mean square error (RMSE) and mean absolute error (MAE) as evaluation indicators to evaluate the experimental results [20, 21]. Given a sparse low-voltage station area data tensor, N represents the total amount of missing data in this tensor, that is, the size of the test set in this experiment. The formula of the evaluation index is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} M_{i,j,k}\left(c_{i,j,k} - \widetilde{c}_{i,j,k}\right)^2}, \tag{19}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} m_{i,j,k}c_{i,j,k} - \widetilde{c}_{i,j,k}, \tag{20}$$

where $c_{i,j,k}$ represents the true value and $\widetilde{c}_{i,j,k}$ represents the predicted value. To make the evaluation more convenient, the mask $m_{i,j,k} \in \{0, 1\}$.

4.2. Experimental Results and Analysis. We consider that in the case of random missing electricity data in low-voltage area and long-term missing data due to faults, this paper is verified by analysing the performance of the proposed method under different time granularities and different training set sizes and the results of long-term missing data completion. Based on the effectiveness of the proposed method and comparing the method in this article with three missing current data complementation methods, it is concluded that the method proposed in this article is better than other methods.

4.2.1. Random Missing Data Completion. To verify the performance of our model, the time interval size was set to 5 minutes, 10 minutes, and 15 minutes, and the experimental results at different time granularities were obtained, as shown in Table 1.

It can be seen from the table that as the time interval increases, the evaluation indicators (RMSE and MAE) of the proposed model show a downward trend, and the complement effect continues to improve. This is because as the time interval increases when predicting missing values, more users' power consumption conditions are taken into account, and the prediction effect will be better.

In addition, we also changed the size of the training set for experiments. Taking $\Delta t = 15$ minutes as an example, we randomly selected 30%, 50%, 70%, and 90% of the original training set as the new training set for experimentation. The experimental results are shown in Table 2. It can be seen from the table that the complete performance of the proposed model increases with the increase of the training set. This is because the larger the training set is, the more the available samples can be used to mine user current data information, thereby obtaining more information and leading to accurate completion results. Therefore, the method in this paper can effectively complete data completion in practice.

4.2.2. Completion of Missing Data throughout the Day. A serious failure may occur during the collection or transmission of the power data by the smart meter, which cannot be recovered in a short time, and the user power data may be lost for a whole day [22]. This paper verifies the effectiveness of the proposed method by randomly discarding current data for several days.

Figure 6(a) shows the actual data and completion results when a user's data are missing for 7 days throughout the day. It can be seen from the figure that the two curves basically overlap, indicating that the completion effect is better. Figure 6(b) is the area chart of the difference between the complement value and the actual value when the entire day is missing for 7 days. It can be seen from the figure that the missing data can still be prepared for the missing data in the entire day.

Table 3 shows the performance data of this method under different missing days. It can be seen from the table that the changes in RMSE and MAE increase with the increase in missing days. When the number of missing days is less than 12 days, both RMSE and MAE are relatively small, indicating that the method proposed in this paper can compensate for the missing days. The entirety is still valid, and the fewer the missing days there are, the better the completion effect. In practice, the missing data will generally be repaired within a week, and the lack of more days rarely happens. Therefore, the method in this article is also more applicable in practice.

4.2.3. Experimental Comparison. The method in this article is compared with three missing current data completion methods (cubic spline interpolation [23], Kalman filtering [24], and tensor completion [25]). Figure 7 shows that the 31-day use of the dataset in the station area is the original complete data. The average absolute error and root mean square error trend are set at 40%–100%. It can be seen from the figure that the completion errors of all methods decrease with the increase of the dataset used. In addition, the accuracy of the station data completion of different methods is also different. The completion error curves (RMSE and MAE) of the three methods of Kalman filtering, cubic interpolation, and tensor completion are all above the joint matrix decomposition completion error curve. That is, when the joint matrix factorization and completion method has the same dataset size, its complete accuracy is higher than that of the other three methods. It is worth noting that the two types of errors are complemented by the joint matrix factorization method; when the dataset size is 50% of the original complete dataset, the error is only approximately equal to 10% of the error of other methods, indicating that the joint matrix factorization method is complementary. Compared with other methods, the whole method is more suitable for the completion of high-deficiency cases.

Figure 8 shows the variation trend of the root mean square error and the mean absolute error of the 31-day data of randomly missing all days in the station area from 1 to 12 days. It can be seen from the figure that the errors of all-day missing data completion of all the completion methods

Table 1: Experimental completion error at the same time granularity.

Time granularity Δt (min)	RMSE/A	MAE/A
5	0.3615	0.2238
10	0.3447	0.2065
15	0.3365	0.1975

Table 2: Experimental completion errors under different training set sizes.

Training set size (%)	RMSE/A	MAE/A
30	0.6615	0.2725
50	0.4647	0.2038
70	0.3065	0.1315
90	0.1005	0.0538

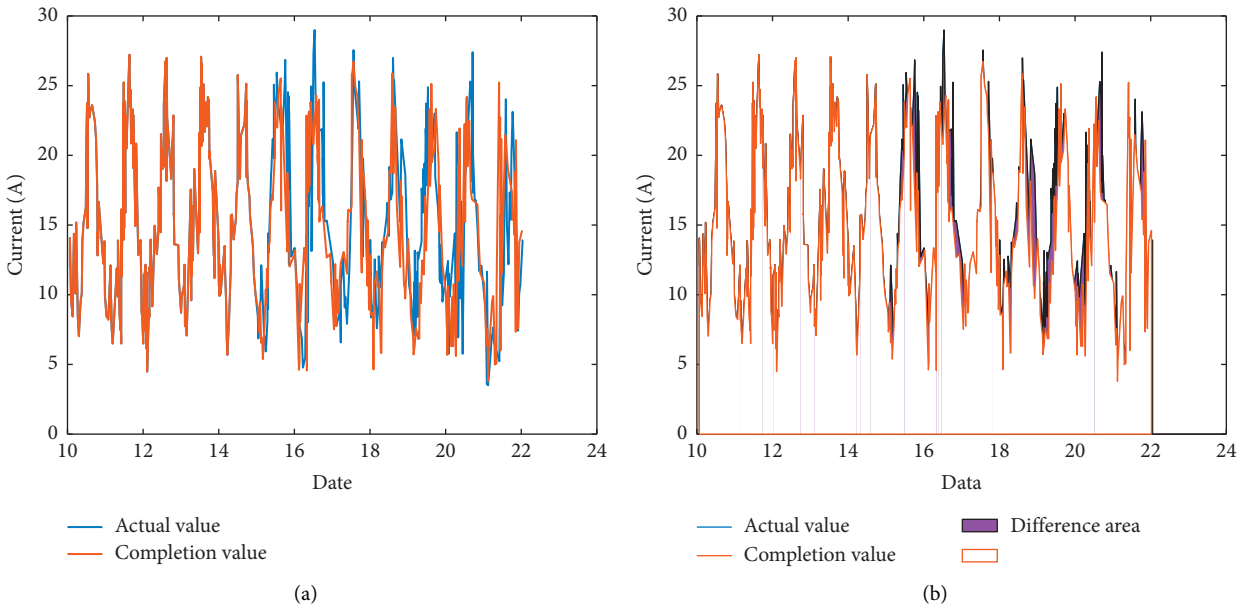

Figure 6: Complete map of the experiment with 7 days missing all day. (a) Completion curve when seven days are missing in the whole day. (b) The area of the difference between the true value and the complement value when the entire day is missing for seven days.

Table 3: Experimental completion error under different missing days.

Missing days/d	RMSE/A	MAE/A
1	0.8615	0.1725
4	1.2047	0.6438
8	1.7765	1.2315
12	2.3605	2.2238
16	4.8725	2.6318
20	6.3215	4.3159
24	7.2638	5.1245

increase with the increase of the number of missing days, and the two error curves of the joint matrix decomposition completion are both below those of cubic interpolation, Kalman filtering, and tensor compensation. The full bottom, that is, the precision, of joint matrix factorization is higher than that of other methods.

Through the comparison, it can be seen that the joint matrix factorization completion method has better

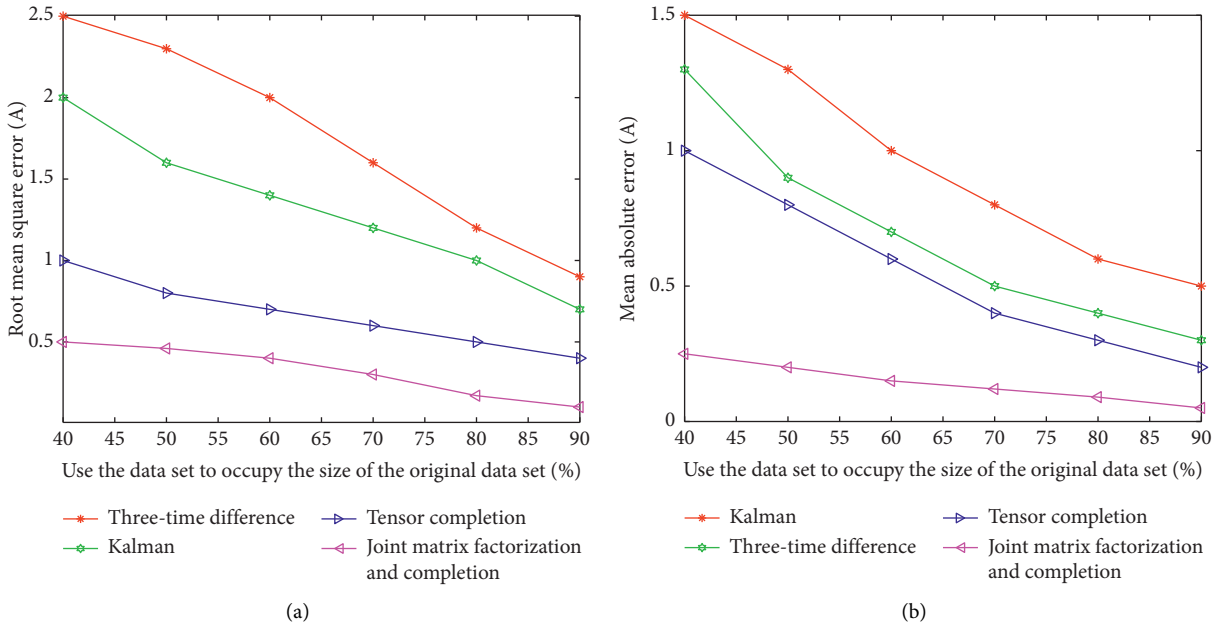

Figure 7: Error comparison results of different dataset sizes. (a) Root mean square error. (b) Mean absolute error.

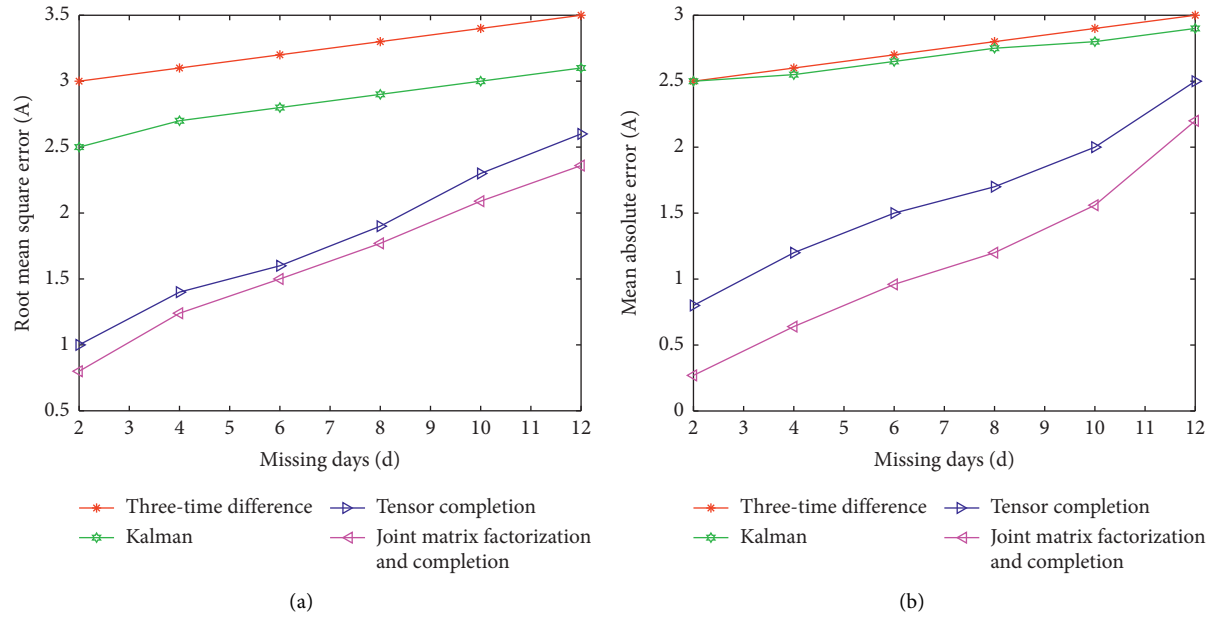

Figure 8: Error comparison results under different missing days. (a) Root mean square error. (b) Mean absolute error.

completion effects than the cubic interpolation, Kalman filter, and tensor completion methods in the case of random missing data or all-day missing data.

4.3. Model Parameter Tuning. In this section, the influence of important parameters on the performance of the model is evaluated, and the experimental results are analysed under different parameter values. In our model, the main focus is on two decomposition modules and local restrictions. Therefore, the main parameters of our research include the number of hidden factors (F), periodic weight parameters (β), and window size ($\mathbf{W}$) and local restriction weight (λ_1).

Figures 9(a) and 9(b) respectively show the changes of RMSE and MAE with the number of latent factors F. RMSE and MAE gradually decrease with the increase of F. This is because the vector in the high-dimensional space can better reflect the influence relationship between different modes. However, when F is too large, the performance of the model begins to degrade. This is because the number of redundant parameters that the model learns is too large due to limited observations, causing the model to overfit. In addition, it can

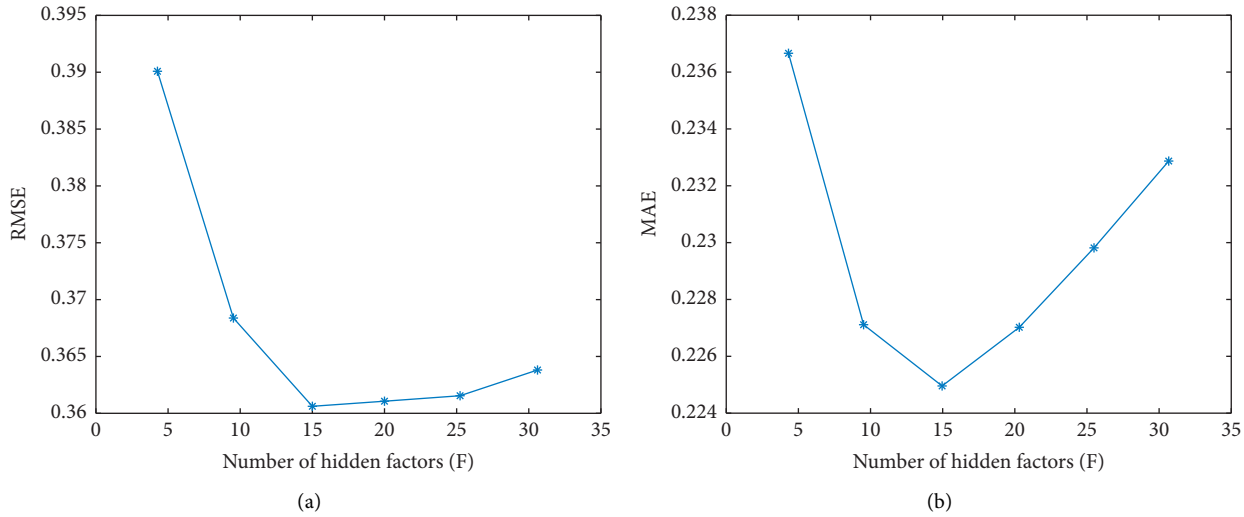

Figure 9: The influence of the number of hidden factors F on the experimental results. (a) RMSE. (b) MAE.

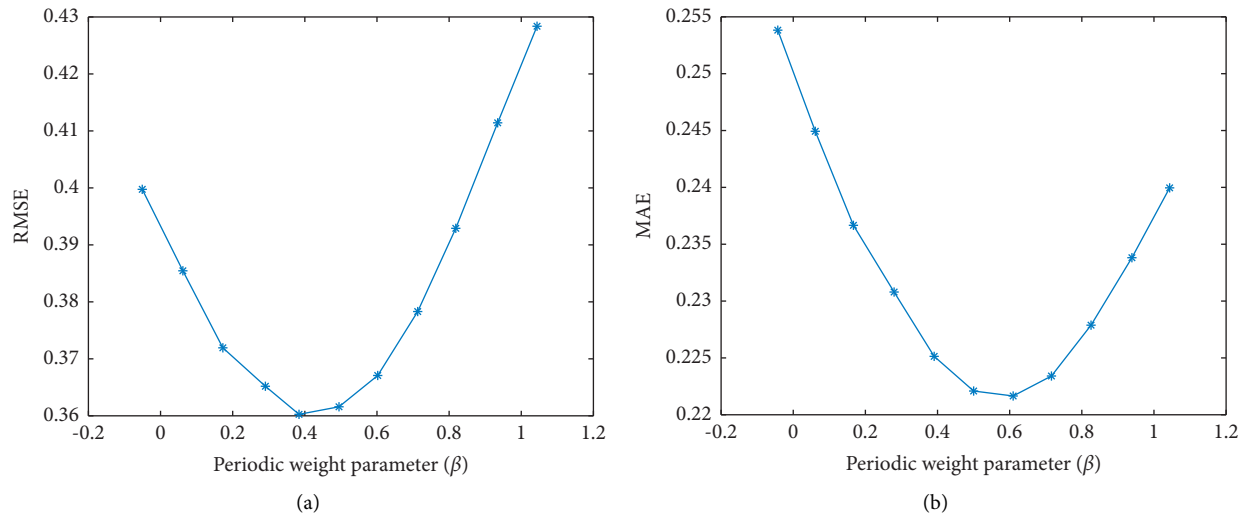

Figure 10: The influence of the weight parameter β on the experimental results. (a) RMSE. (b) MAE.

be seen that the optimal value of RMSE is obtained when F = 15.

In order to explore the influence of the periodic weight parameter β on the model, the value of β is gradually changed from 0 to 1, while the step size is 0.1. The experimental results are shown in Figure 10. A larger value of β contributes to stronger influence of the relationship between the day and the time interval in the model at this time and better performance of the model. Therefore, as β increases, RMSE and MAE gradually become smaller. However, when the value of β exceeds 0.4, the performance of the model begins to gradually decline. This is because the relationship between the day and the time interval is overemphasized in the forecasting process, while the relationship between the user and the time interval is ignored. Here, it is worth noting that $\beta = 0$ means that only periodicity plays a role; when $\beta = 1$, only spatial correlation is used. From Figure 10, we can find that the performance of LPZ can be significantly improved by considering both periodicity and spatial correlation.

In addition, the value of the window size **W** is changed to observe its influence on the model performance. The experimental results are shown in Figure 11. In general, the performance of the model will be improved as **W** increases, because increasing **W** means that more samples can be used to learn temporality. However, RMSE and MAE begin to gradually decrease, when **W** is greater than 4. This phenomenon is caused by the characteristics of timing, that is, the correlation between low-voltage station data at different time intervals will be weakened as the distance increases. Therefore, the time span should be selected appropriately to predict the low-pressure station data at a certain time interval. Too large or too small time granularity of the division will affect the accuracy of the model's prediction results.

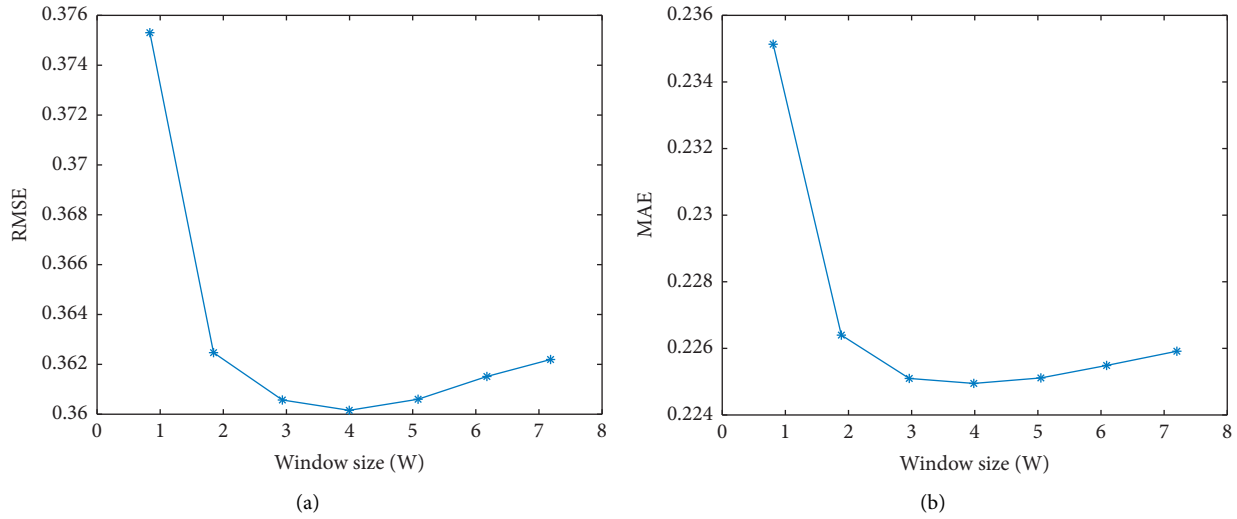

Figure 11: The influence of window size **W** on experimental results. (a) RMSE. (b) MAE.

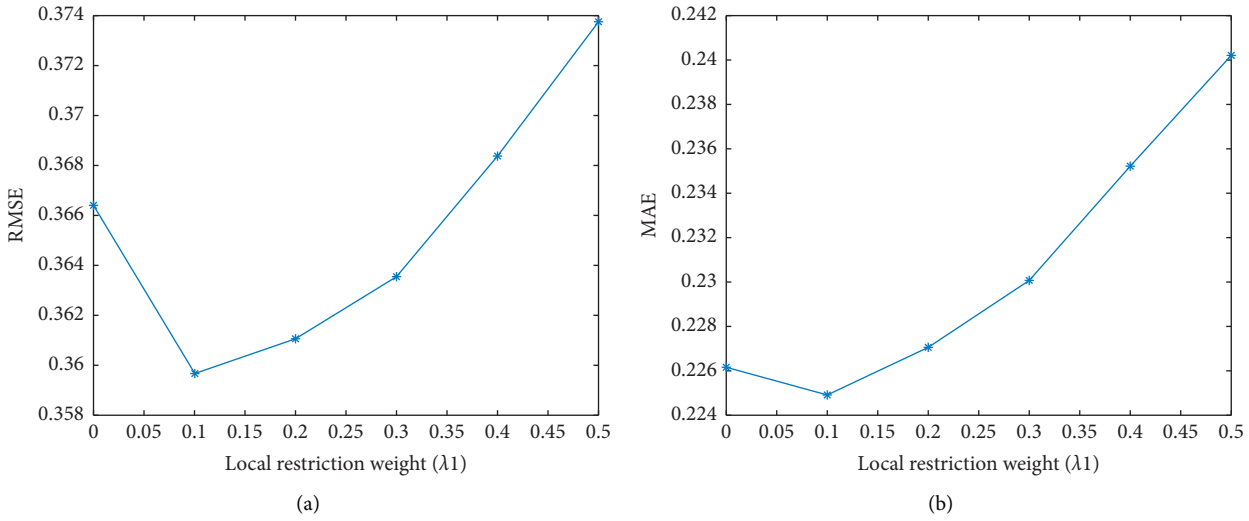

Figure 12: The influence of the local limit parameter λ_1 on the experimental results. (a) RMSE. (b) MAE.

Finally, the performance of the model is analysed under different local restriction weights. As shown in Figure 12, the model performs obviously worse at $\lambda_1 = 0$ than $\lambda_1 = 0.1$, because no local restriction is added to the decomposition module at $\lambda_1 = 0$. This shows that introduction of the time dependency of the relative travel time to the model can improve the performance of the model. When λ_1 is greater than 0.1, the model evaluation index increases with the continuous increase of λ_1. This indicates that we cannot pay too much attention to the timing. It can be seen from Figures 11 and 12 that simultaneous consideration of the periodicity, spatial correlation, and timing of low-voltage station data can improve the performance of LPZ in the task of missing value completion.

5. Conclusions

This paper proposes a model that jointly considers the periodicity, time series, and spatial correlation of the electricity data in the low-voltage station area for the problem of the lack of supplementary power consumption data in the low-voltage station area. The model proposed in this paper can reasonably estimate unknown data based on the observation value of limited low-voltage station area electricity data. Aiming at the characteristics of the electricity data in the low-voltage station area, we fold and stack the electricity data sequence into a three-dimensional tensor form and use it as the input of the joint matrix decomposition module. In the decomposition module, we not only model the

periodicity and spatial correlation but also add local restrictions to make it subject to timing constraints. To verify the effectiveness of the model, we conducted many experiments on a real dataset and compared LPZ with traditional completion methods (cubic interpolation, Kalman filtering, and tensor completion methods). Experiments show that LPZ can not only complement the missing power data but also that the effect of the complement is better than that of traditional complement methods.

To sum up, the completion method based on joint matrix decomposition proposed in this paper is not only suitable to solve the problem of data missing in low-voltage stations but also other data missing problems. With a wide range of application prospects, this method has the advantage of being less affected by the length of the missing data and the location of the missing data. To a certain extent, improving the accuracy of data completion can contribute to higher accuracy of data mining and analysis. Data completion is essential for data mining and analysis and is also of practical value for load forecasting in the power grid and even for power generation forecasting such as photovoltaic power generation and wind power generation.

Data Availability

The experimental data are obtained by Python simulation, and the experimental result diagram is obtained by MATLAB.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] Y. Wadhawan, A. AlMajali, and C. Neuman, "A comprehensive analysis of smart grid systems against cyber-physical attacks," *Electronics*, vol. 7, no. 10, p. 249, 2018.

[2] Y. Zhang, Y. Ting, and M. Guangyu, "Review and prospect of ubiquitous power Internet of things in smart distribution system," *Electric Power Construction*, vol. 40, no. 6, pp. 1–12, 2019.

[3] G. Zhang and J. Guo, "A novel method for hourly electricity demand forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1351–1363, 2020.

[4] S. Oprea, "Data framework for electricity price setting in competitive environment," *Journal of Physics: Conference Series*, vol. 1297, 2019.

[5] M. Zhou, Y. Li, M. J. Tahir, X. Geng, Y. Wang, and W. He, "Integrated statistical test of signal distributions and access point contributions for Wi-Fi indoor localization," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 5, pp. 5057–5070, 2021.

[6] J. Wang, X. Wang, C. Ma, and L. Kou, "A survey on the development status and application prospects of knowledge graph in smart grids," *IET Generation, Transmission & Distribution*, vol. 15, no. 3, pp. 383–407, 2021.

[7] H. Yang, Y. De Ng, and Z. Liu, "Study on electric load forecasting with historical bad data," *Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control*, vol. 45, no. 15, pp. 62–68, 2017.

[8] M. Gurusamy and P. Vijayakumar, "An efficient cloud data center allocation to the source of requests," *Journal of Organizational and End User Computing*, vol. 32, no. 3, pp. 23–36, 2020.

[9] Q. Liu, X. Li, and H. Cao, "Two-dimensional localization: low-rank matrix completion with random sampling in massive MIMO system," *IEEE Systems Journal*, vol. 15, 2020.

[10] P. Gao, M. Wang, and S. G. Ghiocel, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," *IEEE Transactions on Power Systems*, vol. 31, no. 2, pp. 1006–1013, 2015.

[11] M. Yang, Y. Sun, and G. Mu, "Data completing of missing wind power data based on adaptive neuro-fuzzy inference system," *Automation of Electric Power Systems*, vol. 38, no. 19, pp. 16–21, 2014.

[12] G. Tutz and S. Ramzan, "Improved methods for the imputation of missing data by nearest neighbor methods," *Computational Statistics and Data Analysis*, vol. 90, 2015.

[13] H. Gu, T. Wang, Y. Zhu, C. Wang, D. Yang, and L. Huang, "A completion method for missing concrete dam deformation monitoring data pieces," *Journal of Applied Sciences*, vol. 11, no. 1, 2021.

[14] M. Zhou, Y. Long, W. Zhang et al., "Adaptive genetic algorithm-aided neural network with channel state information tensor decomposition for indoor localization," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 5, pp. 913–927, 2021.

[15] L. Li, J. Zhang, and Y. Wang, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2933–2943, 2018.

[16] V. Salerno and G. Rabbeni, "An extreme learning machine approach to effective energy disaggregation," *Electronics*, vol. 7, no. 10, p. 235, 2018.

[17] F. Meng, Q. Ji, H. Zheng, H. Wang, and D. Chu, "Modeling and solution algorithm for optimization integration of express terminal nodes with a joint distribution mode," *Journal of Organizational and End User Computing*, vol. 33, no. 4, pp. 142–166, 2021.

[18] V. Gandikota, D. Kane, and R. K. Maity, "Vqsgd: vector quantized stochastic gradient descent," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 2197–2205, San Diego, CA, USA, April 2021.

[19] H. Duan, X. Xiao, J. Long, and Y. Liu, "Tensor alternating least squares grey model and its application to short-term traffic flows," *Applied Soft Computing*, vol. 89, p. 106145, 2020.

[20] M. Ćalasan, S. H. E. A. Aleem, and A. F. Zobaa, "On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: a novel exact analytical solution based on Lambert W function," *Energy Conversion and Management*, vol. 210, p. 112716, 2020.

[21] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Processing Letters*, vol. 27, pp. 1485–1489, 2020.

[22] A. H. Afsharinejad, C. Ji, and R. Wilcox, "Heterogeneous recovery from large scale power failures," 2020, https://arxiv.org/abs/2012.15420.

[23] S. Huang, J. Tang, and J. Dai, "1DCNN fault diagnosis based on cubic spline interpolation pooling," *Shock and Vibration*, vol. 202013 pages, 2020.

[24] D. Liu, Z. Wang, Y. Liu, and F. E. Alsaadi, "Extended Kalman filtering subject to random transmission delays: dealing with

packet disorders," *Information Fusion*, vol. 60, pp. 80–86, 2020.

[25] C. Cai, G. Li, and H. V. Poor, "Nonconvex low-rank tensor completion from noisy data," 2019, https://arxiv.org/abs/1911.04436.