Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1336900, 11 pages
https://doi.org/10.1155/2021/1336900

Research Article

Missing Data Reconstruction Based on Spectral *k*-Support Norm Minimization for NB-IoT Data

Luo Xuegang,[1] **Lv Junrui**,[1] **and Wang Juan**[2]

[1]*School of Mathematics and Computer Science, Panzhihua University, Panzhihua 617000, China*
[2]*College of Computer Science, China West Normal University, Nanchong 637000, China*

Correspondence should be addressed to Wang Juan; wjuan0712@126.com

Received 23 September 2021; Accepted 16 October 2021; Published 10 November 2021

Academic Editor: Xianyong Li

An effective fraction of data with missing values from various physiochemical sensors in the Internet of Things is still emerging owing to unreliable links and accidental damage. This phenomenon will limit the predicative ability and performance for supporting data analyses by IoT-based platforms. Therefore, it is necessary to exploit a way to reconstruct these lost data with high accuracy. A new data reconstruction method based on spectral *k*-support norm minimization (DR-SKSNM) is proposed for NB-IoT data, and a relative density-based clustering algorithm is embedded into model processing for improving the accuracy of reconstruction. First, sensors are grouped by similar patterns of measurement. A relative density-based clustering, which can effectively identify clusters in data sets with different densities, is applied to separate sensors into different groups. Second, based on the correlations of sensor data and its joint low rank, an algorithm based on the matrix spectral *k*-support norm minimization with automatic weight is developed. Moreover, the alternating direction method of multipliers (ADMM) is used to obtain its optimal solution. Finally, the proposed method is evaluated by using two simulated and real sensor data sources from Panzhihua environmental monitoring station with random missing patterns and consecutive missing patterns. From the simulation results, it is proved that our algorithm performs well, and it can propagate through low-rank characteristics to estimate a large missing region's value.

1. Introduction

An Internet of Things (IoT) platform including sensors, wireless communication, and data processors has been deployed to support remote monitoring and intelligent analysis. At present, a large number of smart industries have emerged, such as smart health, smart campus, smart finance, smart retail, and smart agriculture, which benefit from the rise of the Internet of Things technology. Among them, narrowband IoT (NB-IoT) technology is developing at a rapid speed. Compared with the 3G network, 4G network, wireless sensor network (WSN), Wi-Fi, low-power wide-area network from the LoRa Alliance (LoRaWAN), Zigbee, and so on, it has the characteristics of wide coverage, multiple connections, low rate, low cost, low-power consumption, and excellent architecture [1]. Therefore, NB-IoT is rapidly transforming into a highly heterogeneous ecosystem that provides information exchange among different types of information sensor equipment and communications technologies.

Environmental monitoring platform based on NB-IoT, which provides monitoring, analysis, governance, and protection of environmental quality for relevant government departments, solves the shortcomings of single environmental monitoring, complex wiring, poor transmission capacity, time-consuming, and laborious [2, 3]. IoT has played an important role to strengthen national competitiveness and attracted the attention of industry, academia, government, and regulators worldwide [4]. A large number of IoT platforms have been developed for monitoring, supervision, and management. In addition to collecting data by sensors, IoT platforms also need to have the function of intelligent analysis by big data and artificial intelligence technology. Because these

data are large-scale, high-dimension, and complex structures, there are strict requirements in terms of data integrity, correctness, and on-time delivery. However, in many IoT applications, especially the environmental monitoring of the field environment, it is unavoidable that massive data are lost. On the one hand, owing to unreliable links and accidental damage, data loss or damage may occur during acquisition, transmission, and storage. On the other hand, the limited capability of sensor nodes impacts data collection in terms of energy, storage, and communication. Some issues with battery depletion and device failure result in imperfect data gathering.

Incomplete data collected by sensors make it more difficult to process sensor data, hindering various high-level applications of the Internet of Things, such as intelligent analysis and prediction. If missing data values cannot be accurately recovered, existing intelligent analysis tools cannot be applied. If lost data are deleted directly, a large amount of original data becomes unavailable, which reduces the accuracy and reliability of analysis results and wastes a large amount of data resources. The amount of data collected by the Internet of Things every day around the world is enormous. Currently, that is about 26 billion bytes per day, and the number will increase dramatically. Recovering lost sensor data effectively to analyze IoT applications accurately is a major challenge. Designing effective methods to reconstruct loss sensor data has a wide range of applications. Therefore, it is very urgent and important to design an effective method to reconstruct the missing values in sensor data.

Many methods of populating missing sensor data have been developed by researchers. Methods based on time, space, or a combination of space and time are predominant. However, in some special cases, high-frequency data loss may obscure the temporal and spatial correlation between data. Therefore, it is necessary to study new methods to solve high-frequency data loss. As is known to all, a sensor node usually has multiple sensors integrated. These nodes usually collect multiple monitoring data at the same time, and the data collected by these nodes have a certain correlation. Therefore, it is beneficial to improve the estimation accuracy in predicting sensor data, which can use the correlation between different types of data as a supplement to the internal correlation.

In this paper, a new data reconstruction method based on spectral k-support norm minimization [5] is proposed for NB-IoT data. This method implements the relative density-based clustering algorithm [6] to separate sensors into different groups. The main objective of clustering is to increase the similarity within the same group in terms of the spatial relationship of sensor nodes. Thereafter, we use the correlations of sensor data and its joint low rank; an algorithm based on the matrix weighted Schatten-p norm minimization with automatic weight for filling the multiattribute sensor data within each cluster is developed. Moreover, the alternating direction method of multipliers (ADMM) is used to obtain its optimal solution. Our contributions are summarized as follows:

(1) Considering the uneven distribution of sensor nodes, the relative density-based clustering algorithm is applied to divide the sensor nodes into different clusters according to the distribution around neighbor points of the data points, to ensure that similar patterns of measurement of sensors are within one group

(2) The data reconstruction method based on spectral k-support norm minimization is applied for reconstruction of missing sensor data, taking advantage of the low-rank feature between different attributes of sensor data

The remainder of this paper is organized as follows. In Section 2, background and related work are presented. Section 3 describes our proposed method. The performance of the proposed method is evaluated in Section 4. Section 5 provides our concluding remarks and a simple elaboration for future work.

2. Background and Related Work

NB-IoT, which is a wireless telecommunications technology standard, is developed to connect multiple Internet of Things devices and empower IoT architecture using existing mobile networks. The NB-IoT network has been widely employed in industry, agriculture, healthcare, and logistics, and, quite obviously, in smart cities and buildings. Its purpose is to facilitate the work of technicians, retailers, and operators while handling machines by providing real-time technological data and supporting remote monitoring systems. According to recent studies, by 2026, cellular LPWAN solutions (that is, NB-IoT and LTE-M combined) will be responsible for over 60% of the estimated 3.6 billion low-power wide-area network connections, with the remaining 40% covered by noncellular, of which LoRa and Sigfox will account for the majority. Figure 1 shows NB-IoT end-to-end system architecture including sensor nodes, NB-IoT wireless networks, application server platform, and user terminal equipment. Many techniques in the literature based on temporal methods, spatial methods, and machine learning-based methods have focused on loss data prediction in NB-IoT.

Temporal methods, namely, the mean of observation data [7], the last observation and linear interpolation, and so on, exploit the time correlations between the readings recorded by the same sensor node. Because the data obtained by the sensor node near the monitoring range often varies slowly over a short time period. However, when the observation interval of a given sensor is long or the sensor network changes drastically and irregularly, the temporal methods do not work well, and as the number of consecutive missing readings increases, its effectiveness decreases rapidly.

Spatial methods utilize the spatial correlation between the sensor data of spatially similar sensor nodes at the same monitoring time; generally, the closer the sensor nodes are in space, the greater the correlation of the data is. The missing

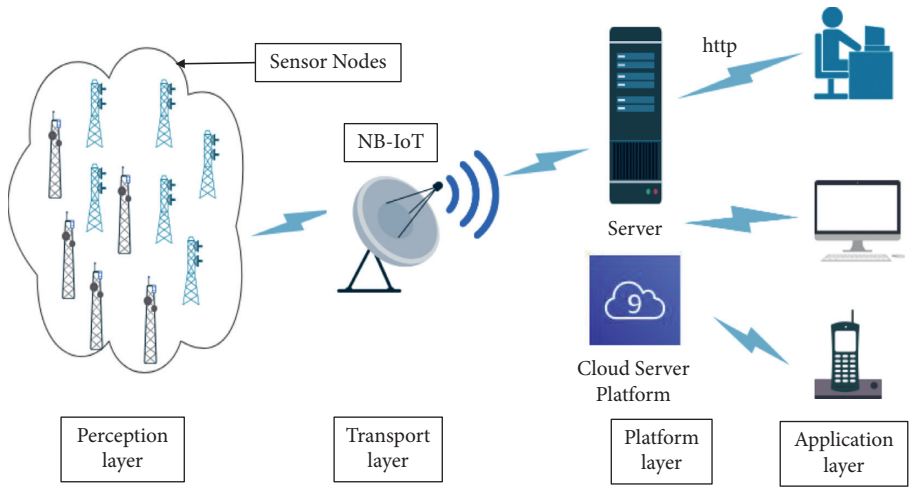


Figure 1: NB-IoT end-to-end system architecture.

data estimation algorithm based on *K*-nearest neighbor [8] takes advantage of the values of the nearest K neighbors to estimate the missing one. Rahman et al. [9] have described a missing data imputation approach that was combining Fourier and lagged k-nearest neighbor for biomedical time series missing data imputation. The nearest neighbor (NN) imputation method that estimates missing data in WSNs by learning spatial and temporal correlations between sensor nodes was proposed by Li and Parker [10]. An MP-BMDI algorithm for univariate time series with large missing gap was proposed by Lee [11]. However, such models, which typically require precise NN distances between sensors, are not suitable only for large missing gap of signals and environments within univariate time series.

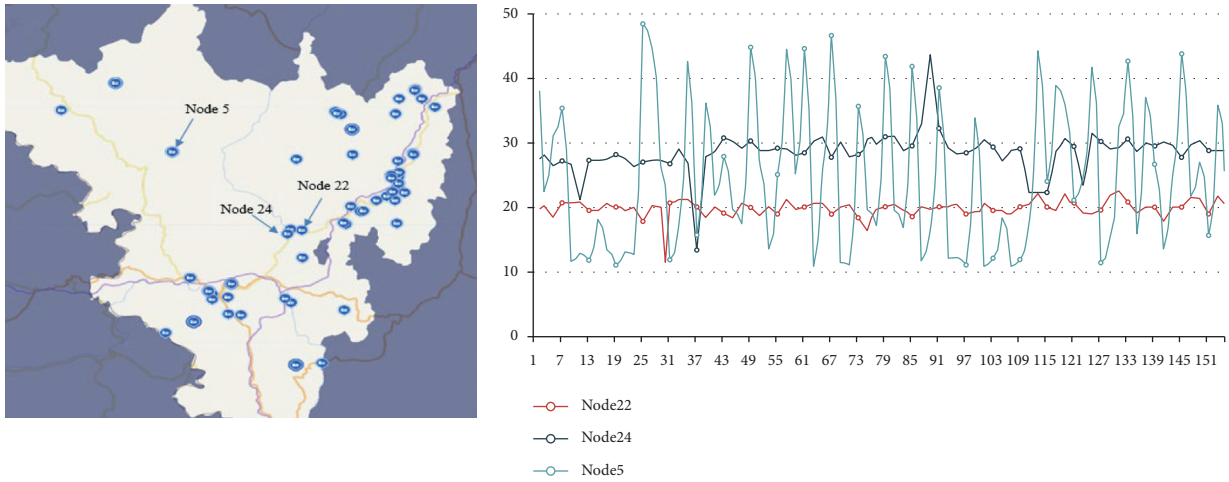
Machine learning-based methods utilizing the machine learning techniques mainly focus on the data correlation structure of the whole dataset to estimate the missing values [12–15]. Here, the multiple linear regression model (MLR) [12] using spatio-temporal correlation was proposed for missing data reconstruction of WSN Data. Zhang and Yin [13] investigated a multivariate time series missing data imputation by recurrent denoising autoencoder and showed legitimate reconstruction performance in a realistically high-dimensional dataset. In [14], the authors presented a novel estimation method of multivariate time series missing data by adapting the bidirectional long short-term memory (LSTM). Pattern sequence forecasting was applied to reconstruct loss data in [15]. The performance of machine learning-based methods depended on accurate historical time series data, but those data are often incomplete and missing.

All the above methods aiming to estimate missing values suffered from over-relying on assumptions about data. For example, when the sampling frequency of the sensor decreases, the usefulness of temporal methods may drop rapidly. Besides, estimation results may be worse as nonexistent spatial correlations are imposed between nearby sensors. In the same way, a priori knowledge that nonexistent temporal correlation was favored for missing data imputation will lead to performance degradation. Therefore, it is necessary to directly exploit sensor latent structures from data without heavily relying on assumptions. Recently, the intrinsic low-rank property of high-dimensional data has been applied to exploit latent structures of sensor data. Weighted nuclear norm minimization [16] was exploited to multiattribute missing data reconstruction for improving IoT data reliability. Despite numerous methods have been developed for imputing missing values, imputation precision and computational complexity are still the major issues for large missing subsequences. To solve this problem, a newly developed algorithm utilized weighted spectral k-support norm minimization that ensures high reconstruction performance for supporting IoT-based applications and data analysis in this paper when large missing subsequences are absent.

3. Our Proposed Method

In this section, the relative density-based clustering applied to separate sensors into different groups is presented; then, a detailed aspect of our proposed reconstruction procedure that tensor composed of sensor multiattribute data is recovered by weighted spectral *k*-support norm minimization is described.

3.1. Sensor Nodes Grouping with the Relative Density-Based Clustering Algorithm. Generally, many sensor nodes are deployed in a monitored region and have spatial correlations from [8]. For example, Figure 2(a) shows the sensor nodes locations from Panzhihua city in China and marked three nodes, among which sensor node 22 and sensor node 24 are close to each other, while sensor node 5 is far away. The deployment locations of sensor nodes are unknown for monitoring purposes. In Figure 2(b), chemical oxygen demand (COD) (unit: mg/L) values of sensor node 5, node 22, and node 24 are curve plotted from the urban environmental water quality monitoring dataset in Panzhihua. Figure 2(b) demonstrates that the trend of monitoring index curve remained consistent of sensor node 5 and sensor node 22.



(a) (b)

FIGURE 2: Sensor nodes with monitoring index location. (a) Sensor nodes locations in a region. (b) COD (unit: mg/L) collected by three sensor nodes.

However, the index curve of sensor node 5 is far from nodes 22 and 24 as monitoring is characterized by concentration and aggregation. Neighboring nodes have similar attributes and similar data changes. Thus, when some of the sensed data of a sensor node are missing, we can estimate them by using the data of the neighboring nodes. This example indicated that not every node in the monitoring region is useful for recovering the missing values. Thus, it is necessary to first divide the sensors into different groups to minimize measurement changes within each group for utilizing similarity well among measured values of neighboring sensors.

The K means algorithm, which is the most typical clustering method [17], can effectively recognize convex clusters. However, K means cannot precisely identify the true clusters for nonconvex clusters. The density-based clustering method can find clusters of various shapes and sizes in data with noisy [6]. The relative density-based clustering algorithm can effectively find clusters in data sets with nonuniform density.

In this work, we apply the relative density-based clustering algorithm to group sensors. This algorithm divides a set of points into many clusters in terms of the relative density of each point and distance so that points in each cluster tend to be close to each other.

Given the data set P and the k-distance of point d, the k -distance neighbors of point d are points whose distance from d is not greater than the k -distance, defined by

$$D_{p\_dist(d)}(d) = \{d \in P \mid dis(b,d) \le k_{\text{distance}}(d)\}. \quad (1)$$

Reachability distance of point d with respect to point s is defined as

$$\text{reach}_{\text{dist}(d,s)} = \max\{k_{\text{distance}}(s) dis(d,s)\}. \quad (2)$$

where $dis(d,s)$ denotes the distance of point d and point s and $k_{\text{distance}}(s)$ is the k-nearest neighbors.

The local reachability density of d is the inverse of the average based on the k-nearest neighborhood of a point d, defined as

$$\text{reach\_dsity}_k(d) = \frac{1}{\left(\sum_{s \in D_k(d)} \text{reach}_{\text{dist}(d,s)}\right) / \left|D_k(d)\right|}. \quad (3)$$

We simplify the notation and use $D_k(d)$ as a shorthand for $D_{p\_dist(d)}(d)$.

Then, the relative density of d with respect to its $D_{p\_dist(d)}(d)$ neighbors denoted as $\text{rel\_dsity}_k(d)$ is defined as

$$\text{rel\_dsity}_k(d) = \frac{\sum_{s \in D_k(d)} \text{reach\_dsity}_{k(s)} / \text{reach\_dsity}_{k(d)}}{\left|D_k(d)\right|}. \quad (4)$$

The main steps of the relative density-based clustering algorithm are as follows:

Step 1: set data set of points P, and initialize the clustering center node-set $C = \{0\}$

Step 2: calculate the distance between point j and point r, and then obtain dist[j, r]

Step 3: for each point, calculate the k-distance neighbors and calculate the k-distance neighbors, reachability distance, local reachability density, and relative density according to (1)–(4)

Step 4: select points whose relative density is less than 1 and store them in set IC, and C is filled with assignment method (IC, dist[j, r]) (see [6])

Step 5: for each point d in data set P, group them to the nearest cluster C_i

Step 6: iterate steps 2, 3, 4, and 5 until set C no longer changes or the difference between adjacent iterations is smaller than the given threshold ε

After the above algorithm, sensors of the urban environmental water quality monitoring dataset in Panzhihua are divided into five groups (see Figure 3).

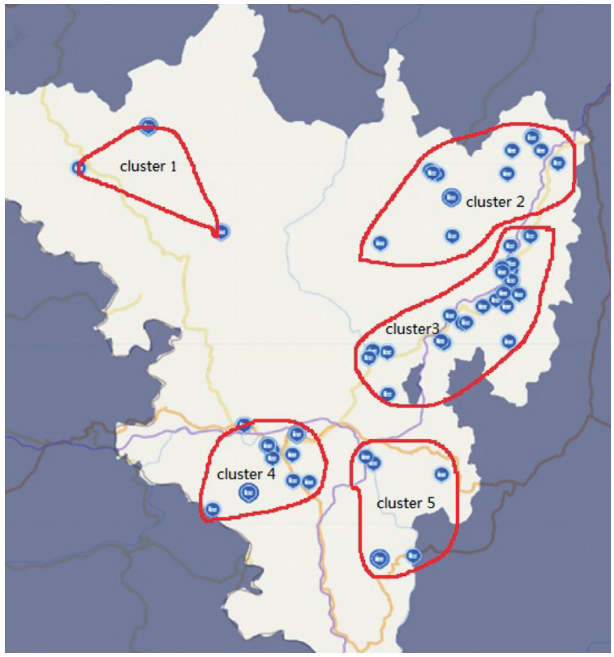


Figure 3: Clustering result of sensor location in the Panzhihua region.

3.2. Data Reconstruction with Spectral k-Support Norm Minimization

3.2.1. Spectral k-Support Norm. As sensor data have a low-rank characteristic, the problem of missing data reconstruction can be regarded as the rank minimization problem. The goal is to recover missing values of a matrix $M \in \mathbb{R}_r^{m\times n}$ from partial observations, supposing that its rank $r \ll \min\{m,n\}$ among all matrices with the same observations, and m and n are the number of rows and columns of the matrix. The following model for the solution of the rank minimization problem can estimate the unknown M:

$$\min_{\mathbf{X}} \operatorname{rank}(X), s.t. \|X - M\|_{\mathrm{F}}^2 < \delta, X_{ij} = M_{ij}, \quad \forall (i,j) \in \Omega, \tag{5}$$

where Ω is the set of the observed indices; $\delta > 0$ is a small constant; and $\|.\|^{\mathrm{F}}$ represents the norm. It is difficult to solve problem (5), that is, NP-hard [18]. To obtain a solution in polynomial time, the nuclear norm is proposed as a substitute. The nuclear norm-based matrix completion model is formulated as follows:

$$\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_*, s.t. \|X - M\|_{\mathrm{F}}^2 < \delta, X_{ij} = M_{ij}, \quad \forall (i,j) \in \Omega, \tag{6}$$

where $\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X)$ is the kernel norm, λ is a constant and greater than zero, and $\sigma_i$is the ith singular value. Problem (6) adopts singular value threshold contraction of the observation matrix, namely,

$$\begin{cases} M = U\Sigma V^T, \\ X = U\zeta_\lambda(\Sigma)V^T, \end{cases} \tag{7}$$

where $U\Sigma V^T$ is SVD decomposition, $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{r\times n}$ are the decomposed orthogonal matrix, $\zeta_\lambda(\Sigma)_{ii} = \max(\Sigma_{ii} - \lambda, 0)$ is the soft threshold function of a diagonal matrix Σ, and Σ_{ii} is the iterative solution of diagonal matrix elements in (6) until convergence.

Recently, the spectral k-support norm [19] has been shown to have good estimation properties in low-rank matrix learning problems and better captures the spectral decay of the underlying model, and the spectral k-support norm-based matrix completion model is as follows:

$$\min_{\mathbf{X}} \|X - M\|_{\mathrm{F}}^2 + \|\mathbf{X}\|_{(k)}^{msp} \text{ s.t. } X_{ij} = M_{ij}, \quad \forall (i,j) \in \Omega. \tag{8}$$

where $\|X\|_{(k)}^{msp}$denotes spectral k-support norm. The spectral k-support norm is the gauge function whose unit ball is the convex set:

$$\operatorname{conv}\{\sigma | \|\sigma\|_0 \le k, \|\sigma\|_F \le 1\}, k \in \mathbb{N}_+. \tag{9}$$

Thus, the unit ball of the spectral k-support norm is the convex hull of matrices, and its rank is no larger than k and Schatten-p norm no larger than 1. The spectral k-support norm-based model has shown superior performance over the nuclear norm-based models.

It can be the following explicit computation:

$$\|X\|_{(k)}^{msp} := \|\sigma\|_{(k)}^{msp} = \left[\sum_{i=1}^{k-l-1} (\sigma_i)^2 + \frac{1}{l+1}\left(\sum_{j=k-l}^{D} \sigma_j\right)^2\right]^{1/2}, \tag{10}$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{\min(m,n)}$ are the singular values of Min nonincreasing order and l is the unique integer in $[0, k-1]$ satisfying

$$\sigma_{k-l-1} \ge \frac{1}{l+1} \sum_{j=k-l}^{\min(m,n)} \sigma_j \ge \sigma_{k-l}. \tag{11}$$

3.2.2. Data Reconstruction Algorithm and ADMM-Based Solution. The sensor data gathered in one node can be organized by the following format of sensor ID standing for sensor identity number, timestamp representing sampling time, and the monitoring index parameters including chemical oxygen demand, ammonia nitrogen, residual chlorine, and humidity.

Let $\mathcal{T} \in \mathbb{R}^{Q\times P\times C}$ be a tensor, whose data are composed of Q monitoring index parameters collected by P nodes within C time slots. The entry $\mathcal{T}$ is represented by $t_{q,p,c}$. Because of data loss in NB-IoT, $\mathcal{T}$ is usually an incomplete tensor. A set of entries of $\mathcal{T}$ were intact information gathered by sensors. We use $[\mathcal{G}_\Omega(\cdot)]$ to indicate the sampling operator, which is defined as follows:

$$[\mathcal{G}_\Omega(\mathcal{T})]_{q,p,c} = \begin{cases} t_{q,p,c}, \text{if } (q,p,c) \in \Omega, \\ 0, \quad \text{otherwise.} \end{cases} \tag{12}$$

The sensor missing data reconstruction problem with spectral k-support norm minimization is defined as follows:

$$\begin{aligned} &\min \|\mathscr{L}\|_{(k)}^{tsp} + \lambda \|\mathscr{R}\|_1 \\ &s.t. \quad \mathscr{G}_\Omega(\mathscr{T}) = \mathscr{G}_\Omega(\mathscr{L}) \text{ and } \mathscr{H} * (\mathscr{T} - \mathscr{L}) = \mathscr{R}, \end{aligned} \tag{13}$$

where $\|\bullet\|_1$ is the Frobenius norm, $\|\cdot\|_{(k)}^{tsp}$ denotes tensor spectral k-support norm, that is, an extension of matrix spectral k-Support norm $\|\cdot\|_{(k)}^{msp}$, and $\mathscr{H}$ is an identity mapping and the element-wise projection, correspondingly. As the sensor data tensor is low rank, the recovery tensor $\mathscr{L}$ is restricted by tensor spectral k-support norm minimization in (13).

The alternate direction multiplier method (ADMM) algorithm, a convex optimization algorithm, is a good candidate for the low-rank model problem when compared with the other state-of-the-art splitting proximal algorithms. Therefore, in the present study, we attempted to use the ADMM algorithm to solve the formulated problem. To apply the ADMM method to solve equation (13) in the paper via ADMM-type algorithm, adding a new tensor-valued auxiliary variable $\mathfrak{J}$, the augmented Lagrangian is given by

$$\begin{aligned} \mathscr{D}_\beta(\mathscr{L}, \mathscr{R}, \mathfrak{J}) = &\frac{1}{2}\|\mathscr{L}\|_{(k)}^{tsp} + \lambda \|\mathscr{R}\|_1 \\ &+ \langle \mathfrak{J}, \mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R} \rangle \\ &+ \frac{\beta}{2}\|\mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R}\|_F^2, \end{aligned} \tag{14}$$

where β is a penalty parameter from the augmented Lagrangian formulation and λ is a constant from ℓ_1 norm. In the following, we carry out the alternative update to the variables $\mathscr{L}_t$ and $\mathscr{R}_t$ at iteration t in detail.

(1) Update $\mathscr{L}_t$ as follows:

$$\begin{aligned} \mathscr{L}_t = \arg\min_{\mathscr{L}} &\frac{1}{2}\|\mathscr{L}\|_{(k)}^{tsp} + \langle \mathfrak{J}_{t-1}, \mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R}_{t-1} \rangle \\ &+ \frac{\beta}{2}\|\mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R}_{t-1}\|_F^2. \end{aligned} \tag{15}$$

To separate $\mathscr{H}$ apart from $\mathscr{L}$, the preconditioned ADMM approximates $1/2\|\mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R}_{t-1}\|_F^2$ with second-order Taylor expansion around $\mathscr{L}_{t-1}$ as

$$\frac{1}{2}\|\mathscr{H} * (\mathscr{T} - \mathscr{L}_{t-1}) - \mathscr{R}_{t-1}\|_F^2 - \langle \mathscr{H}^{\mathrm{T}} * (\mathscr{H} * (\mathscr{T} - \mathscr{L}_{t-1}) - \mathscr{R}_{t-1})\mathscr{L} - \mathscr{L}_{t-1} \rangle + \frac{\delta}{2}\|\mathscr{L} - \mathscr{L}_{t-1}\|_F^2. \tag{16}$$

Incorporating equation (16) into equation (15) gives

$$\begin{aligned} \mathscr{L}_t = \arg\min_{\mathscr{L}} &\frac{1}{2\beta\delta}\|\mathscr{L}\|_{(k)}^{tsp} + \langle \mathfrak{J}_{t-1}, \mathscr{H} * (\mathscr{T} - \mathscr{L}) - \mathscr{R}_{t-1} \rangle \\ &+ \frac{1}{2}\left\| \mathscr{L} - \left(\mathscr{L}_{t-1} + \mathscr{H}^T * (\mathscr{H} * (\mathscr{T} - \mathscr{L}_{t-1}) - \mathscr{R}_{t-1}) + \frac{\mathfrak{J}_{t-1}}{\beta\delta} \right) \right\|_F^2. \end{aligned} \tag{17}$$

Thus, according to the proximal operator for TSP-k norm in [5], problem (17) has a closed-form solution introduced to solve this problem. The computation of the proximal map in equation (17) has been given by

$$\mathscr{L}_t = \mathrm{Prox}_{\frac{1}{2\beta\delta}\|\cdot\|_{tsp,k}^2}\left(\mathscr{L}_{t-1} + \mathscr{H}^T * (\mathscr{H} * (\mathscr{T} - \mathscr{L}_{t-1}) - \mathscr{R}_{t-1}) + \frac{\mathfrak{J}_{t-1}}{\beta\delta} \right). \tag{18}$$

(2) Update $\mathscr{R}_t$ as follows:

$$\begin{aligned}\mathscr{R}_t &= \arg\min_{\mathscr{R}} \lambda\|\mathscr{R}\|_1 + \langle \mathfrak{Y}_{t-1}, \mathscr{H} * (\mathscr{T} - \mathscr{L}_t) - \mathscr{R}\rangle + \frac{\beta}{2}\|\mathscr{H} * (\mathscr{T} - \mathscr{L}_t) - \mathscr{R}\|_F^2 \\ &= \arg\min_{\mathscr{R}} \frac{\lambda}{\beta}\|\mathscr{R}\|_1 + \frac{1}{2}\left\|\mathscr{R} - \mathscr{H} * (\mathscr{T} - \mathscr{L}_t) - \frac{\mathfrak{Y}_{t-1}}{\beta}\right\|_F^2.\end{aligned} \tag{19}$$

The problem of equation (19) can be used as the proximal operator of the ℓ_1 norm to solve

$$\mathscr{R}_t = \mathrm{Prox}_{\frac{\lambda}{\beta}\|\cdot\|_1}\left(\mathscr{H} * (\mathscr{T} - \mathscr{L}_t) + \frac{\mathfrak{Y}_{t-1}}{\beta}\right). \tag{20}$$

Equation (20) can be efficiently computed by the element-wise soft thresholding operation and is as follows:

$$\mathrm{Prox}_{\frac{\lambda}{\beta}\|\cdot\|_1}\left(\mathscr{F}_{i.j.k}\right) = \mathrm{sign}\left(\mathscr{F}_{i.j.k}\right)\max\left\{\left|\mathscr{F}_{i.j.k}\right| - \frac{\lambda}{\beta}, 0\right\}, \tag{21}$$

where

(3) Update $\mathfrak{Y}_t$ as follows:

$$\mathfrak{Y}_t = \mathfrak{Y}_{t-1} + \beta(\mathscr{H} * (\mathscr{T} - \mathscr{L}_t) - \mathscr{R}_t). \tag{22}$$

(4) *Stopping Criterion*. Given a tolerance $\varepsilon > 0$, we check the termination condition of primal variables $\|\mathscr{L}_t - \mathscr{L}_{t-1}\|/\|\mathscr{T}\| \le \varepsilon$.

After discussing the appearing subproblem, the complete DR-SKSNM algorithm for sensor data reconstruction in NB-IoT is presented in Algorithm 1.

The computational complexity of DR-SKSNM consists mainly of two parts: one is the complexity of the relative density-based clustering algorithm and the other is the cost of tensor reconstruction computation. When the relative density-based clustering algorithm is used to separate sensor nodes, c cluster centers are first set randomly, and then computing the distance from n points to c centers, the distance between the single node and cluster center is calculated as P, and finally the above process is repeated for t times, so the complexity of relative density-based clustering is O(cnpt). Usually, c, P, and t can be regarded as constants; therefore, the computational cost of relative density-based clustering can be simplified to be linear, namely, O(n). From Algorithm 1, the sensor data reconstruction algorithm mainly involves computation of $\mathscr{L}_t$,$\mathscr{R}_t$, and $\mathfrak{Y}_t$, the parameters iterated T times, so the computational complexity of the matrix completion algorithm is O(nT). Thus, the computational complexity of DR-SKSNM in this paper is O(nT).

4. Experimental Results

In this section, the performance of our proposed algorithm called DR-WSKSNM is investigated by comparing it over other alternative benchmark algorithms.

4.1. Data Description. The water quality monitoring NB-IoT platform in Sichuan Province (China) from industrial and urban sewage discharge monitoring stations was utilized to collect original data. The monitoring variables including chemical oxygen demand (COD) (unit:mg/L), ammonia nitrogen (unit:mg/L), total nitrogen (unit:mg/L), and pH are measured and stored in Sichuan Province intelligent ecological management platform, every 10 minutes from January 1, 2020, to the present. The data from January 1, 2020 (00 : 00 : 00 a.m.) to August 30, 2020 (23 : 59 : 59 p.m.) are particularly extracted to utilize for analyzing reconstruction performance. Table 1 shows some analyzing samples from sensor ID 105 of the water quality monitoring NB-IoT platform.

To validate the availability of the algorithm objectively, two types such as random missing pattern and consecutive missing pattern are applied to verify [11].

(i) *Random Missing Pattern (RMP)*. This pattern is suitable for missing data with a random time and random sensor to be missing. α represents the data missing rate.

(ii) *Sequence Missing Pattern (SMP)*. This pattern reflects that all data are missed after a certain sampling time point owing to running out of batteries or experiencing a loss of connectivity to the acquisition platform. Therefore, we randomly selected 5% of nodes as objective nodes that suffer from sequence data missing. Then, the missing subsequences with different time intervals such as one-hour, three-hour, six-hour, and nine-hour missing are utilized to assess the validity of the proposed method.

Input: T, H;
Initialize: $L_0 = T, L_1 = 0, R_0 = 0, S_0 = 0$ and set parameters λ, β, δ, ε, $t = 1 \ldots T$.
While $\|L_t - L_{t-1}\|/\|T\| \geq \varepsilon$, then
  $t = t + 1$;
  Update L_t by equation (17);
  Update R_t by equation (20);
  Update S_t by equation (21);
End while
Output: L_{t-1}.

Algorithm 1: DR-SKSNM algorithm for NB-IoT sensor data reconstruction.

Table 1: Data samples from the water quality monitoring NB-IoT platform.

Monitoring time	COD (mg/L)	Ammonia nitrogen (mg/L)	Total nitrogen (mg/L)	pH
2020-01-01 00 : 00 : 00	12.009	2.419	9.21	7.02
2020-01-01 00 : 10 : 00	12.27	2.124	9.27	7
2020-01-01 00 : 20 : 00	12.239	2.07	9.28	6.98
2020-01-01 00 : 30 : 00	11.81	1.923	9.41	6.95
...	...	...	...	...

4.2. Reconstruction Performance Indicators. Our algorithm is evaluated with the existing data reconstruction methods based on the following two metrics.

4.2.1. MAPE (Mean Absolute Percentage Error). MAPE (p, q) is defined as the average of absolute percentage errors of forecasts, which is measured by the following:

$$\mathrm{MAPE}(p, q) = \frac{100}{w} \sum_{i=1}^{w} \left| \frac{p_i - q_i}{q_i} \right|, \tag{23}$$

where w denotes the length of missing subsequence, P is a predicted value, and q indicates a true value. A smaller $\mathrm{MAPE}(p, q)$ signifies a better approach for the reconstruction of missing values.

4.2.2. RMSE (Root Mean Square Error). RMSE (P, q) is a metric for measuring reconstruction error and is denoted as the average squared difference between the predicted data (P) and the true data (q). A better performance method will have a lower value of RMSE. It is defined as follows:

$$\mathrm{RMSE}(p, q) = \sqrt{\frac{1}{w} \sum_{i=1}^{w} (p_i - q_i)^2}. \tag{24}$$

In our experiment, we set some parameters empirically [20], specifically, $\lambda = 0.4$, $\beta = 0.5$, $\delta = 0.01$, and $\varepsilon = 10^{-3}$. The parameter ε affects the speed of convergence of the ADMM algorithm. All simulations were run in the MATLAB 2016b environment on a laptop equipped with an Apple M1 processor and 8 GB RAM.

4.3. Performance Comparison and Discussion. To examine the performance of our proposed algorithm, four different methods based on tensor data reconstruction are compared with the true values for validation. These true values are set as ground truth to make quantitative comparisons by utilizing two metrics, namely, MAPE and RMSE.

Figure 3 shows the locations of sensor nodes and their clusters in the Panzhihua region from China. Here, the number of clusters is 5. The corresponding sensor nodes of each cluster are included within the red solid lines in Figure 3. Our DR-SKSNM algorithm for missing data reconstruction was applied within each group except cluster 5 due to fewer sensor nodes in it. Tensors are composed of attributes of monitor index, nodes, and time slots in each cluster. In our experiment, the outputs of the recovered sensor data with that of existing algorithms, namely, DRAWNNM [16], HaLRTC [20], and ADMAR [21] were compared with the output of the proposed method. By comparison, the correct rank (3, 2, 3) and a higher rank are set to execute Tucker decomposition in random missing patterns and consecutive missing patterns.

Tables 2 and 3 demonstrate, respectively, the quantitative results of four methods on indicators for the task of reconstructing missing values with missing rate of RMP and missing gap of SMP. The best indicator values are labeled as italic and bold in Tables 2 and 3. From the comparison, the performance of our algorithm is better than that of the existing algorithms from two metrics on the COD and total nitrogen datasets. When the missing rate is less than 20%, the DRAWNNM algorithm performs as well as the other methods. Moreover, when the missing rate is higher than 20%, our indicator values are the best. When comparing the performance of DR-SKSNM on two different datasets, better performance in terms of MAPE and RMSE appears on the COD dataset than on the total nitrogen dataset. This is because the COD dataset has preferable monitoring data that are more similar among measured values of neighboring sensors. The result of total nitrogen dataset is particularly unstable because it is based on the influence of parameter drifting on sampling data.

Table 2: Performance evaluation for four methods by average MAPE and RMSE on the COD and total nitrogen datasets (unit: mg/L) with RMP.

Dataset	Missing rate	DRAWNNM		HaLRTC		ADMAR		Proposed method	
		MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
COD	$\alpha = 10\%$	15.745	38.475	16.701	39.421	16.945	39.584	14.546	38.491
	$\alpha = 20\%$	42.608	85.270	43.624	88.130	43.123	94.220	40.964	85.560
	$\alpha = 40\%$	99.641	189.678	91.248	195.781	90.478	198.657	88.735	188.953
Total nitrogen	$\alpha = 10\%$	6.175	45.285	7.930	48.698	7.270	52.357	6.190	45.025
	$\alpha = 20\%$	27.457	89.324	28.907	93.234	29.098	94.876	26.896	87.324
	$\alpha = 40\%$	86.323	175.239	88.680	168.853	89.510	158.680	80.764	155.381

Table 3: Performance evaluation for four methods by average MAPE and RMSE on the COD and total nitrogen datasets (unit: mg/L) with SMP.

Data type	Missing gap	DRAWNNM		HaLRTC		ADMAR		Proposed method	
		MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
COD	One-hour	15.704	27.360	17.693	26.030	16.942	19.660	15.067	23.140
	Three-hour	32.439	93.060	33.641	83.780	32.380	92.730	30.290	70.570
	Six-hour	72.110	139.570	89.820	141.330	99.550	147.910	64.980	137.640
	Nine-hour	131.983	265.522	126.638	253.591	131.628	280.032	119.190	245.730
Total nitrogen	One-hour	23.589	29.452	25.159	32.021	24.929	37.230	23.976	33.381
	Three-hour	45.178	96.271	48.627	93.737	46.680	94.560	40.221	89.321
	Six-hour	82.275	169.601	86.921	163.260	90.172	177.290	84.217	157.290
	Nine-hour	156.210	274.289	147.210	283.240	158.258	288.126	149.274	265.216

The corresponding results for each dataset are illustrated in Figures 4–7. First, on the COD dataset, Figure 4 demonstrates the visual comparison of reconstruction value curves of four methods and real values curve at position node 26 in cluster 4 from Figure 3 with a missing gap of six-hour. As noticed in Figure 4, the curve shape of reconstructed values resulted from DR-SKSNM only corresponds more closely to that of real values. This substantiates that DR-SKSNM yields superior performance than other compared methods when sensors in one group have similar patterns of measurement. The reconstructed values from DRAWNNM, HaLRTC, and ADMAR also match less with real values than DR-SKSNM in Figures 4–6 though they do generate the up-down trends.

As illustrated in Figures 5 and 6, our algorithm merely demonstrates its ability with missing gaps of nine-hour and twelve-hour for ammonia nitrogen and total nitrogen datasets at position node 36 in cluster 5 and node 21 in cluster 3 from Figure 3, respectively. Our proposed method has better consistency between reconstructed data and real data from Figures 5 and 6. Because the reconstructed data curves of compared methods fluctuate greatly, there are many deviations from the real data. In particular, it is not difficult to find the curve trend from Figure 6 that the local error of reconstructed data of the ADMAR method is high. Notably, when the data missing gaps are high, the error rate of the proposed algorithm is considerably lower than that of the other three algorithms. The reason is that our algorithm DR-SKSNM uses the relative density-based clustering algorithm to group sensors, which greatly enhances the connection of sensor data and improves the accuracy of the algorithm. As shown in Figure 7, a visual comparison of the

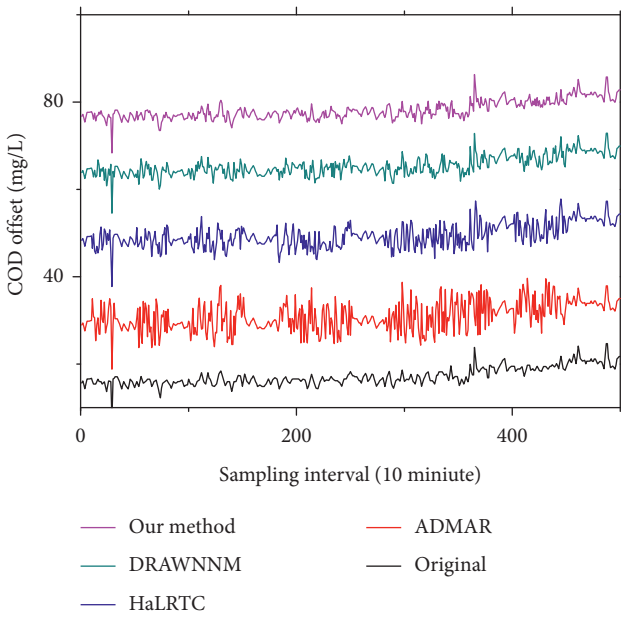


Figure 4: Visual comparison of reconstruction curves of four methods and original data curve (monitoring variable dataset: COD).

reconstruction curves of four approaches and the original data curve for pH datasets with missing gaps of two hours at location node 26 from Figure 3 is reported. From Figure 7, it can be seen that the DR-SKSNM algorithm also has a good performance and can reconstruct the missing data well. Moreover, the DR-SKSNM algorithm is superior to the other algorithms, including DRAWNNM.

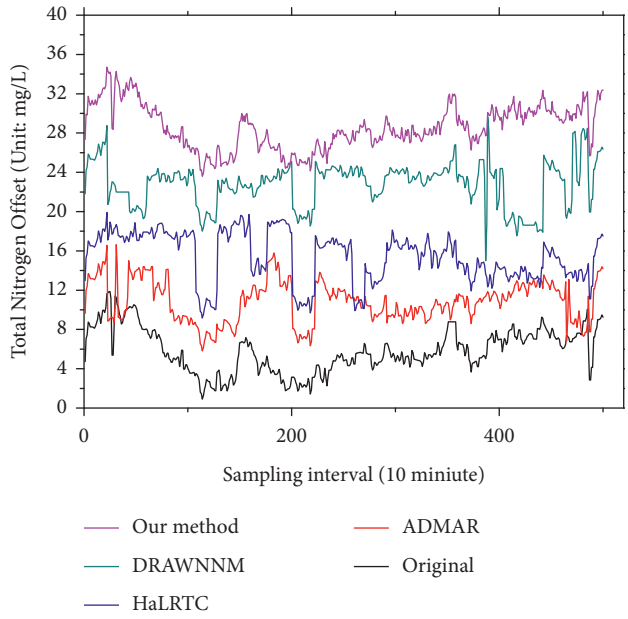


Figure 5: Visual comparison of reconstruction curves of total four methods and original data curve (monitoring variable dataset: total nitrogen).

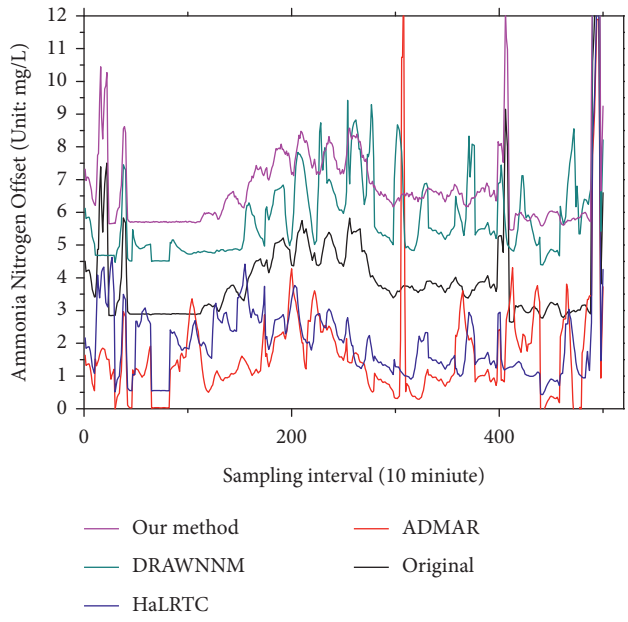


Figure 6: Visual comparison of reconstruction curves of total four methods and original data curve (monitoring variable dataset: ammonia nitrogen).

Consequently, the reconstruction performances of all the existing methods including the proposed algorithm decisively vary on the characteristics of each dataset for testing. Thus, the existing techniques are not as effective as our proposed algorithm, and we further confirm that DR-SKSNM is a better option for reconstruction missing data in real-life NB-IoT applications.

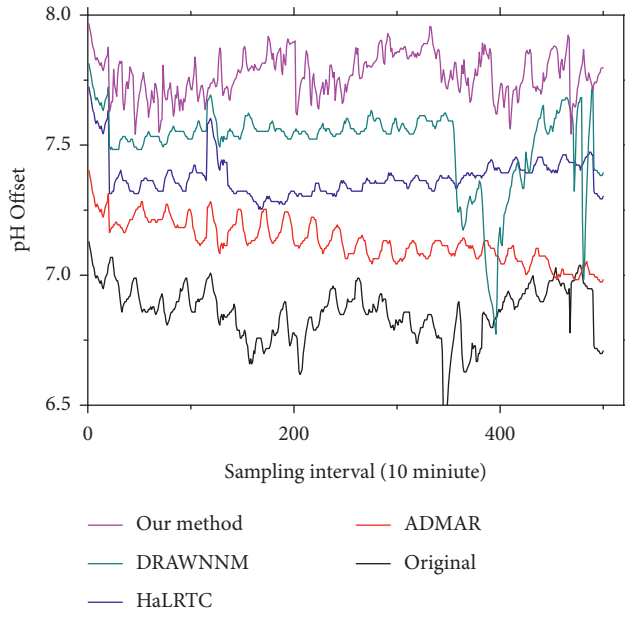


Figure 7: Visual comparison of reconstruction curves of four methods and original data curve (monitoring variable dataset: pH).

5. Conclusion

Missing data from sensors hinder many NB-IoT applications. To solve the problem, it is necessary to estimate the missing data as accurately as possible. In this paper, we presented the DRAWNNM algorithm as a novel approach to reconstruct missing data for NB-IoT applications. This algorithm has been tested using four different real datasets currently stored in our databases. We examined the accuracy of reconstructed values generated from DR-SKSNM, by comparing with three existing methods such as DRAWNNM, HaLRTC, and ADMAR, using two performance metrics (MAPE and RMSE). Through the experiments, we demonstrate that our proposed algorithm outperforms these existing methods when dealing with high data missing rates and large data missing gaps as validated both by performance in terms of MAPE and RMSE. Overall, this work provides strong evidence that DRAWNNM is potentially superior to compared algorithms in terms of accuracy and computational complexity. Finally, the intriguing future work in this paper is that the correlation among more multiple various sensors is taken into account and machine learning techniques are used to identify a superior subsequence.

Data Availability

The data used to support the findings of this study were supplied by luoxuegang under license and so cannot be made freely available. Requests for access to these data should be made to corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Innovation Foundation of Sichuan Province of China under Grant 2019088, and in part by the Science and Technology Support Project of Panzhihua City of Sichuan Province of China under Grant 2021CY-S-6.

References

[1] M. Ganzha, M. Paprzycki, and W. Pawlowski, "Semantic interoperability in the Internet of Things; an overview from the INTER-IoT perspective," *Journal of Network and Computer Applications*, vol. 81, no. MAR., pp. 111–124, 2016.

[2] M. V. Ramesh, K. V. Nibi, and A. Kurup, "Water quality monitoring and waste management using IoT," in *Proceedings of the 2017 IEEE Global Humanitarian Technology Conference (GHTC)*, IEEE, San Jose, CA, USA, October 2017.

[3] L. Teng, X. Min, and J. Chen, "Automated water quality survey and evaluation using an IoT platform with mobile sensor nodes," *Sensors*, vol. 17, no. 8, p. 1735, 2017.

[4] S. Shanzhi Chen, H. Hui Xu, D. Dake Liu, B. Hu, and H. Wang, "A vision of IoT: applications, challenges, and opportunities with China perspective," *IEEE Internet of Things Journal*, vol. 1, no. 4, pp. 349–359, 2014.

[5] J. Lou and Y.-M. Cheung, "Robust low-rank tensor minimization via a new tensor spectral k -support norm," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2314–2327, 2020.

[6] Y. Wang and Y. Yang, "Relative density-based clustering algorithm for identifying diverse density clusters effectively," *Neural Computing & Applications*, vol. 33, no. 16, pp. 10141–10157, 2021.

[7] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "TinyDB: an acquisitional query processing system for sensor networks," *ACM Transactions on Database Systems*, vol. 30, no. 1, pp. 122–173, 2005.

[8] L. Pan and J. Li, "K-nearest neighbor based missing data estimation algorithm in wireless sensor networks," *Wireless Sensor Network*, vol. 2, no. 2, pp. 115–122, 2010.

[9] S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg, "Combining Fourier and lagged k -nearest neighbor imputation for biomedical time series data," *Journal of Biomedical Informatics*, vol. 58, pp. 198–207, 2015.

[10] Y. Li and L. E. Parker, "Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks," *Information Fusion*, vol. 15, pp. 64–79, 2014.

[11] G. Ho Lee, J. Han, and J. K. Choi, "MPdist-based missing data imputation for supporting big data analyses in IoT-based applications," *Future Generation Computer Systems*, vol. 125, pp. 421–432, 2021.

[12] Y. Zaid, B. Zhang, W. M. Ismael, Y. Xie, G. N. Surname, and H. Wang, "A spatio-temporal multiple linear regression missing data reconstruction approach for improving wsn data reliability," in *Proceedings of the 2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pp. 1–6, July 2021.

[13] J. Zhang and P. Yin, "Multivariate time series missing data imputation using recurrent denoising autoencoder," in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 760–764, IEEE, San Diego, CA, USA, November 2019.

[14] P. Thi-Thu-Hong, "Machine learning for univariate time series imputation," in *Proceedings of the 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, IEEE, Hanoi, Viet Nam, October 2020.

[15] N. Bokde, M. W. Beck, F. Martínez Álvarez, and K. Kulat, "A novel imputation methodology for time series based on pattern sequence forecasting," *Pattern Recognition Letters*, vol. 116, pp. 88–96, 2018.

[16] X. Yu, X. Fan, K. Chen, and S. Duan, "Multi-attribute missing data reconstruction based on adaptive weighted nuclear norm minimization in IoT," *IEEE Access*, vol. 6, pp. 61419–61431, 2018.

[17] A. K. Jain, "Data clustering: 50 Years beyond K-means," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008*, W. Daelemans, B. Goethals, and K. Morik, Eds., vol. 5211, Springer, Berlin, Heidelberg, 2008, Lecture Notes in Computer Science.

[18] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[19] A. M. Mcdonald, M. Pontil, and D. Stamos, "Fitting spectral decay with the k-support norm," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (PMLR)*, vol. 51, pp. 1061–1069, Cadiz, Spain, May 2016.

[20] Y. Shao, Z. Chen, F. Li, and C. Fu, "Reconstruction of big sensor data," in *Proceedings of the 2nd IEEE International Conference Computer Communication (ICCC)*, pp. 1–6, Chengdu, China, October 2016.

[21] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.