Data Loss and Reconstruction in Wireless Sensor Networks

Linghe Kong, Member, IEEE, Mingyuan Xia, Xiao-Yang Liu, Guangshuo Chen, Yu Gu, IEEE Member, Min-You Wu, IEEE Senior Member, Xue Liu, IEEE Member

Abstract—Reconstructing the environment by sensory data is a fundamental operation for understanding the physical world in depth. A lot of basic scientific work (e.g., nature discovery, organic evolution) heavily relies on the accuracy of environment reconstruction. However, data loss in wireless sensor networks is common and has its special patterns due to noise, collision, unreliable link, and unexpected damage, which greatly reduces the reconstruction accuracy. Existing interpolation methods do not consider these patterns and thus fail to provide a satisfactory accuracy when the missing data rate becomes large. To address this problem, this paper proposes a novel approach based on compressive sensing to reconstruct the massive missing data. Firstly, we analyze the real sensory data from Intel Indoor, GreenOrbs, and Ocean Sense projects. They all exhibit the features of low-rank structure, spatial similarity, temporal stability and multi-attribute correlation. Motivated by these observations, we then develop an environmental space time improved compressive sensing (ESTI-CS) algorithm with a multi-attribute assistant (MAA) component for data reconstruction. Finally, extensive simulation results on real sensory datasets show that the proposed approach significantly outperforms existing solutions in terms of reconstruction accuracy.

Index Terms—Wireless Sensor Networks, Data Loss and Reconstruction, Compressive Sensing

1 INTRODUCTION

People investigate the environment in order to understand our physical world. Recently, wireless sensor networks (WSNs) [15], [20] are widely adopted to gather sensory data and reconstruct the environment in the cyber space [11]. There are plenty of environment monitoring applications under the water [30], in the forest [23], and on the volcano [28]. An environment matrix (EM) is a common way to represent a dynamic environment. An EM is an \( n \times t \) matrix that records data from \( n \) sensors over \( t \) time intervals. Environment reconstruction [13] attempts to obtain the full and accurate EM from raw sensory data.

Motivation: A great deal of basic scientific work heavily depends on the accuracy of environment reconstruction. For example, scientists reveal the nature of ocean currents from accurate underwater temperature data [30], understand the demand for plant evolution based on the light condition in the forest [23], and discover the eruption omen by monitoring the shake of the volcano [28].

However, since data gathering is largely affected by hardware and wireless conditions, a raw dataset usually has notable missing data. Furthermore, missing data become larger as WSNs grow in scale [3]. Consequently, data loss becomes the key challenge against accurate reconstruction. It is urgent and important to design effective methods to recover incomplete EMs.

Existing approaches and limitations: The missing value problem is fundamental in dataset field. Lots of work has been contributed such as K-Nearest Neighbors (KNN) [7], Delaunay Triangulation (DT) [13], and Multi-channel Singular Spectrum Analysis (MSSA) [32]. These methods are often used when there are only a few missing values, but cannot be applied when the missing data grow.

Compressive Sensing (CS) [5], [8] is a powerful and generic technique for estimating missing data. CS can recover an entire dataset from only a small fraction of data as long as these data contain sparse/low-rank features. So far, CS has been utilized to reconstruct network traffic [31], refine localization [24] and improve urban traffic sensing [19]. However, since a WSN has unique data loss patterns, directly applying CS on EM interpolation cannot gain satisfactory accuracy.

Our work and contribution: Our work is fourfold: Firstly, we analyze real environmental data from Intel Indoor [1], GreenOrbs [23], and OceanSense [30] projects. We confirm the massive data loss in general applications and mine the specific data loss patterns in WSNs. Then we reveal four features in environmental datasets: 1) Low-rank structure. A complete EM can be represented by a few principle data, which underpins the applicability of CS. 2) Time stability. The sensory values of one certain node are usually similar at adjacent time slots. 3) Space similarity. The sensory values of neighbor nodes are similar. 4) Multi-
attribute correlation. Multiple environmental attributes have strong correlation in some cases. For example, the change tendency of temperature and light are similar in the OceanSense project [30].

Secondly, motivated by these observations, we design a novel environmental space time improved compressive sensing (ESTI-CS) algorithm for estimating the missing data. ESTI-CS embeds customized features into the baseline CS to deal with the specific data loss patterns, which computes the minimal low-rank approximations of the incomplete EM and refines the interpolation with spatio-temporal features.

Thirdly, when multiple attributes from the same dataset have strong correlation, we design a multi-attribute assistant MAA component to leverage this feature for better reconstruction accuracy.

Fourthly, we evaluate the effectiveness of our approach based on trace-driven simulations. We demonstrate that ESTI-CS can outperform existing approaches such as CS, KNN, DT, and MSSA when the raw data contain diverse real loss patterns. Typically, ESTI-CS can achieve an effective environment reconstruction with less than 20% error when there are 90% missing data in the collected data. In addition, MAA further enhances the performance of ESTI-CS in extensive simulations.

2 RELATED WORK
The missing value problem is common in datasets [3]. A great deal of existing work has devoted to interpolating the missing data. K-Nearest-Neighbor (KNN) [7] is a classical local interpolation method. KNN simply utilizes the values of the nearest K neighbors to estimate the missing one. It is frequently used in many low-fidelity estimation cases. Delaunay Triangulation (DT) [13] is a typical global refinement method, which treats the gathered data as vertices. DT takes advantage of these vertices and their global errors to build virtual triangles for data interpolation. It is widely adopted in computer vision for surface rendering. Multi-channel Singular Spectrum Analysis (MSSA) [32] is a data adaptive and nonparametric method based on the embedded lag-covariance matrix. MSSA is often used in geographic data recovery.

Despite much progress in the area of data interpolation, existing methods are suitable for only few missing values, but perform poorly when the loss rate grows high, which is common in WSN cases.

Compressive sensing (CS) is an advanced method to recover the whole condition with just a few sampled data [6], [8]. CS-based methods have been developed for network traffic estimation [31], road traffic interpolation [19], and localization in mobile networks [24]. CS has also witnessed wide applications in WSNs, e.g., recovering signals under noisy background [2], balancing load via compressive data gathering [22]. However, the study of CS for environment reconstruction in WSNs is still vacant.

Existing CS-based interpolation methods cannot be directly applied for accurate environment reconstruction due to two reasons: 1) CS-based methods require the dataset to have inherent structures. Features that are extracted from network traces [31] or road traffic [19] are not applicable for WSN data. 2) CS theory performs well when the missing values follow the Gaussian or pure random distribution [18], [27]. However, as shown in Section 4.2, the loss patterns of WSNs do not satisfy these prerequisites.

To address the above challenges, an effective environment reconstruction approach in WSN is required to deal with the massive data loss problem as well as to study the WSN-specific loss patterns.

3 PROBLEM FORMULATION
3.1 Environmental Data Reconstruction
Rebuilding the virtual environment (such as the dynamic temperature) in cyber space based on the sensory data is called environment reconstruction.

In environment reconstruction systems, sensor nodes are scattered in the given area. Suppose totally n sensor nodes are deployed. The monitoring period includes t time slots. Each sensor node reports its sensory data once per time slot through wireless transmission. x(i, j) denotes the sensory data of node i at time slot j, where i = 1, 2 · · · n and j = 1, 2 · · · t.

Definition 1 Environment Matrix (EM): is a mathematical method to describe the dynamic environment. An EM is defined by X = (x(i, j))n×t.

A complete EM represents that every data in the matrix are successfully collected, i.e., no data loss.

Definition 2 Binary Index Matrix (BIM): is an n × t matrix, which indicates if the data points at the corresponding positions in an EM are missing. BIM is defined as:

\[ B = (b(i, j))_{n \times t} = \begin{cases} 0 & \text{if } x(i, j) \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases} \] (1)

Definition 3 Sensory Matrix (SM): is an n×t matrix, which records the raw data collected from WSNs. Due to the presence of missing data, elements of SM are either x(i, j) gathered by WSNs or zero (missing data point).

Thereby, an SM is an incomplete EM. An SM is denoted by S and can be presented by:

\[ S = B \circ X. \] (2)

3.2 Problem Statement
Data reconstruction is to rebuild the real environment (EM) based on the gathered sensory data (SM).

Definition 4 Reconstructed Matrix (RM): is generated by interpolating the missing values in an SM to approximate EM. RM is denoted by \( \hat{X} = (\hat{x}(i, j))_{n \times t} \).

1. In this paper, AB presents the matrix production of A and B. A \( \circ B \) presents the element-wise production of A and B.
Problem: Environment Reconstruction in Sensor Network (ERSN): Given an SM $S$, ERSN problem is to find an optimal RM $X$ that approximates the original EM $X$ as closely as possible. i.e.,

$$\text{Objective: } \min ||X - \hat{X}||_F,$$

Subject to: $S$.

where $|| \cdot ||_F$ is the Frobenius norm used to measure the error between matrix $X$ and $\hat{X}$. For calculating, take $X$ as an example, $||X||_F = \sqrt{\sum_{i,j}(x(i,j))^2}$.

In ERSN problem, the objective is to minimize the absolute error. In order to measure the reconstruction error in different scenarios among different methods, we further define the following metric.

**Definition 5 Error Ratio (ER):** is the metric for measuring the reconstruction error after interpolation:

$$\epsilon = \sqrt{\sum_{i,j:b(i,j)=0}(x(i,j) - \hat{x}(i,j))^2} \div \sqrt{\sum_{i,j:b(i,j)=0}(x(i,j))^2}.$$

Note that the condition $b(i,j) = 0$ in Eq. (4) indicates that only errors on the missing data are counted.

## 4 DATA LOSS IN SENSOR NETWORKS

In this section, we analyze the data loss in real WSN datasets. The three datasets are from Intel indoor [1], GreenOrbs [23], and OceanSense [30] projects.

### 4.1 Massive Data Loss

Through statistics analysis, we verify that the significant data loss exists in all of these original datasets.

We investigate totally 54 nodes and 84,600 time slots (one month) data from the Intel Indoor dataset, where 23% data points are missing. The GreenOrbs dataset also observes 35% data loss. And this loss is even larger in OceanSense, which is about 64% for 20 nodes and 5,040 time slots (one week). We find that the data loss is common and significant in real WSNs.

### 4.2 Data Loss Pattern

Traditional work usually assumes that the data loss follows a random distribution [19], [32]. However, this claim is not correct in WSN applications. In terms of the nature of WSNs, we synthesize several typical data loss patterns.

**Pattern 1 Element Random Loss (ERL):** This is the simplest loss pattern. Data in the matrix are dropped independently and randomly. Missing data points are randomly distributed in a SM. The noise and collision [12] in WSNs are the root causes of this pattern.

**Pattern 2 Block Random Loss (BRL):** Data from adjacent nodes in adjacent time slots are dropped together. Congestion [10] always leads to data loss on high-density sensor nodes during a period of time.

**Pattern 3 Element Frequent Loss in Row (EFLR):** Unreliable links [29] are common phenomenon in real wireless scenarios. When the link quality is not good, sensory data are prone to loss due to the intermittent transmission. In EFLR, elements in some particular rows have a higher missing probability.

**Pattern 4 Successive Elements Loss in Row (SELR):** A certain node starts losing from a particular time slot. This type of loss occurs when some sensor nodes are damaged or run out of energy [26].

**Pattern 5 Combinational Loss (CL):** In real world, data loss is a combination of loss patterns above.

## 5 ENVIRONMENTAL DATA MINING

### 5.1 Ground Truth

In order to discover the environmental features, the complete datasets are needed as the ground truth. However, EMs from the three original datasets cannot be directly utilized since they all have considerable data loss. To generate applicable EMs, we perform preprocessing on the raw datasets, which selects the small but complete subsets from these three datasets. The size and time interval of selected matrices are shown in Table 1. As a result, six EMs are generated from preprocessing: indoor temperature, indoor light, forest temperature, forest light, ocean temperature, and ocean light.

### 5.2 Low-Rank Structure Discovery

Environmental data of different locations over different times are not independent. There exists inherent structure or redundancy. We mine these features in above selected datasets by Singular Value Decomposition (SVD), which is an effective non-parametric technique for revealing the hidden structure [16].

Any $n \times t$ matrix $X$ can be decomposed into three matrices according to SVD:

$$X = U\Sigma V^T = \sum_{i=1}^{\min(n,t)} \sigma_i u_i v_i^T,$$

where $V^T$ is the transpose of $V$, $U$ is an $n \times n$ unitary matrix (i.e., $UU^T = U^TU = I_{n \times n}$), $V$ is a $t \times t$ unitary matrix (i.e., $VV^T = V^TV = I_{t \times t}$), and $\Sigma$ is an $n \times t$ diagonal matrix constraining the singular values $\sigma_i$ of $X$. Typically, the singular values in $\Sigma$ are sorted, i.e., $\sigma_i \geq \sigma_{i+1}$, $i = 1, 2, \cdots \min(n,t)$, where $\min(n,t)$ is the number of singular values. The rank of a matrix, denoted by $r$, is equal to the number of its non-zero singular values. If $r < \min(n,t)$, the matrix is considered as low-rank.
In Eq. (5), the singular value $\sigma_i$ also indicates the energy of the $i$-th principal component. The total energy is equal to the sum of all singular value $\sum_{i=1}^{\min(n,t)} \sigma_i$. According to PCA, a low-rank matrix [31] exhibits that its first $r$ singular values occupy the total or near-total energy $\sum_{i=1}^{r} \sigma_i \approx \sum_{i=1}^{\min(n,t)} \sigma_i$.

In Fig. 1, we illustrate the distribution of singular values in 6 EMs. The X-axis presents the $i$-th singular values, which is normalized by $\min(n,t)$ because the scales of 6 EMs are different. The Y-axis presents the values of the sum of first $i$-th singular value, which is normalized by $\max(\sigma_i)$ due to the same reason. This figure suggests that the energy is always contributed by the top several singular values in real environments. For example, the top 5% singular values contribute all energy in Indoor-Temp. The universal existence of $\sum_{i=1}^{r} \sigma_i \approx \sum_{i=1}^{\min(n,t)} \sigma_i$ and $r << \min(n,t)$ reveals that EMs exhibit obvious low-rank structures. Low-rank features [19] serve for the prerequisite for using compressive sensing.

(Refer to the supplemental file for the mining of temporal stability and spatial similarity features.)

### 5.3 Multi-Attribute Correlation

We are aware of the following two facts in real WSN applications. (1) Usually, WSNs gather multiple attributes simultaneously, e.g., a TelosB node [21] senses three environmental attributes: temperature, light and humidity. (2) Multiple attributes have correlations in some applications. For instance, the empirical study [17] reveals that several attributes do have relationship such as relative humidity and dewpoint temperature. Therefore, we propose to mine and exploit such correlations to further optimize the accuracy of environment reconstruction.

Joint Sparse Decomposition: In order to mine the correlations, a Joint Sparse Decomposition (JSD) method is proposed to jointly divide multi-attribute EMs into a public sub-matrix $W$ and multiple private sub-matrices $\Delta$. All sub-matrices have the same size of EMs, but their magnitudes are smaller. Suppose two attributes $X_1 = (x^{(1)}_1, \ldots, x^{(t)}_1)$ and $X_2 = (x^{(1)}_2, \ldots, x^{(t)}_2)$, where $x^{(j)}_k$ presents the $j$-th column vector of EM $X_k$, $j = 1, 2, \ldots, t$. For both column vector $x^{(j)}_1$ and $x^{(j)}_2$, the goal is to split them into:

$$\begin{align*}
x^{(j)}_1 &= w^{(j)} + \delta^{(j)}_1, \\
x^{(j)}_2 &= w^{(j)} + \delta^{(j)}_2, \\
w^{(j)} &= \Psi v^{(j)},
\end{align*}$$

where $w^{(j)}$ is the public sub-vector of $x^{(j)}_1$ and $x^{(j)}_2$, which is the multiplication of a wavelet basis $\Psi$ [4] and a sparse vector $v^{(j)}$ satisfying $w^{(j)} = \Psi v^{(j)}$. The private sub-vectors are represented by $\delta^{(j)}_1$ and $\delta^{(j)}_2$ respectively. According to Compressive Sensing theory [8][5], $v^{(j)}$, $\delta^{(j)}_1$, and $\delta^{(j)}_2$ are obtained by solving an $l_1$-norm minimization problem as the following:

$$\hat{\vartheta} = \text{argmin} ||\theta||_1, \quad s.t. \quad x = A\vartheta.$$  

(7)

where $|| \cdot ||_1$ is the $l_1$-norm, $\vartheta = (v^{(j)}T, \delta^{(j)}_1T, \delta^{(j)}_2T)T$, $x = (x^{(j)}_1T, x^{(j)}_2T)T$, and $A = (\Psi, I, 0; \Psi, 0, I)$. It was proved in [8] that solving Eq.(7) is NP-hard, so we adopt the least angle regression method, which was proposed in [9], to obtain $\hat{\vartheta}$. Then the public sub-vector $w^{(j)}$, the private sub-vectors $\delta^{(j)}_1$, and $\delta^{(j)}_2$ can be calculated from $\hat{\vartheta}$.

Applying JSD onto every column vector, $X_1$ and $X_2$ are decomposed as

$$\begin{align*}
X_1 &= W + \Delta_1, \\
X_2 &= W + \Delta_2.
\end{align*}$$

(8)

Correlation: The energy fraction of public sub-matrix $W$ is used to measure the correlation between two attributes, where the total energy is the sum of all singular values of three sub-matrices $W$, $\Delta_1$, and $\Delta_2$. Fig.2 shows the energy fraction of sub-matrices after JSD in diverse groups. Group (a) shows the results of JSD on two irrelevant random matrices. The public sub-matrix $W$ contains only 7% of total energy. Group (b) shows the results of indoor-temp and indoor-light. The change of indoor-light has mutations by manually switching lights on/off, which leads to low correlation with indoor-temp $W = 11%$. The results of forest-temp and forest-light are shown in group (c). Both outdoor light and temperature vary according to the sun. However, due to the influence of tree shade, the correlation is not very strong. So $W$ contains 29% of total energy, while the private sub-matrices $\Delta_1$ and $\Delta_2$ contain 35% and 36% respectively. And sensor nodes are fully exposed under the sun in...
We propose a novel missing data estimation approach to address ERSN problem. The proposed algorithm, namely environmental space time improved compressive sensing (ESTI-CS), takes into consideration the spatio-temporal features to optimize the estimation accuracy.

6.1 Compressive Sensing based Approach

Since we have revealed the low-rank structure in most real environment datasets, we propose to use CS method to estimate missing data from the SM.

The goal of solving ERSN problem is to estimate \( \hat{X} \). According to Eq. (5), any matrix can be decomposed by SVD into \( \sum_{i=1}^{\text{rank}} \sigma_i u_i v_i^T \). Through the inverse process, we can also create an \( r \)-rank approximation \( \tilde{X} \) by using only the \( r \) largest singular values and abandoning the others:

\[
\sum_{i=1}^{r} \sigma_i u_i v_i^T = \tilde{X}.
\]  

This \( \tilde{X} \) is known as the best \( r \)-rank approximation that minimizes the error measured by Frobenius norm. Nevertheless, the optimal \( \hat{X} \) cannot be obtained directly by this way as we do not know matrix \( X \) and the proper rank in advance.

Thus we propose to find \( \hat{X} \) as follows:

Objective: \( \min(\text{rank}(\hat{X})) \),
Subject to: \( B \circ \hat{X} = S \).  

We make this assumption according to two reasons. On the one hand, since the reconstructed matrix (R-M) is generated from the sensory matrix (SM), it is reasonable to be as close as SM. On the other hand, like the environmental matrix (EM), RM should also have a low-rank structure. Given this, it is still difficult to solve this minimization problem because it is non-convex. To bypass this difficulty, we take advantage of the SVD-like factorization, which re-writes Eq. (5) as:

\[
\hat{X} = U \Sigma V^T = LR^T,
\]  

where \( L = U \Sigma_{1,2} \) and \( R = V \Sigma_{1,2} \). Substituting Eq. (11) to Eq. (10), we can solve the minimization problem according to the compressive sensing theory in [5], [8]. Specifically, if the restricted isometry property holds [25], minimizing the nuclear norm can result to rank minimization exactly for a low-rank matrix. Hereby, we just need to find matrix \( L \) and \( R \) that minimize the summation of their Frobenius norms:

Objective: \( \min(||L||_F^2 + ||R||_F^2) \),
Subject to: \( B \circ (LR^T) = S \).  

Looking for \( L \) and \( R \) that strictly satisfy Eq. (12) is likely to fail due to two reasons. First, EMs usually approximate low-rank but not exact low-rank. Second, noises in sensory data may lead to the over-fitting problem if strict satisfaction is required. Thus, instead of solving Eq. (12) directly, we solve the following equation using the Lagrange multiplier method:

\[
\min(||B \circ (LR^T) - S||_F^2 + \lambda(||L||_F^2 + ||R||_F^2)),
\]  

where the Lagrange multiplier \( \lambda \) allows a tunable tradeoff between rank minimization and accuracy fitness. This solution provides the low-rank approximation but not strict satisfaction.

In Eq. (13), 1) \( B \) and \( S \) are known, 2) any \( || \cdot ||_F \) is non-negative, 3) the optimal values approximate 0 by minimizing all non-negative parts. Hence, \( L \) and \( R \) can be estimated in this optimization problem.

6.2 Environmental Spatio-Temporal Improvement

ESTI-CS includes two key components: On the other hand, after exploiting the temporal stability and spatial similarity features, we complete ESTI-CS approach by developing Eq. (13) as following:

\[
\min(||B \circ (LR^T) - S||_F^2 + \lambda(||L||_F^2 + ||R||_F^2) + ||BLR^T T||_F^2)
\]  

\[
+ ||L||_F^2 + ||R||_F^2 + ||T||_F^2)
\]
where \( \mathbb{H} \) and \( T \) are the spatial and temporal constraint matrices respectively. Three subjects \( \| \mathbb{H} L R^T \|_F^2 \), \( \| L R^T T \|_F^2 \), and \( \| B \circ (L R^T) - S \|_F^2 \) are set to be the same order of magnitude, whose coefficients are 1. Otherwise, they may overshadow the others when solving Eq. (14).

**Temporal stability improvement**: The temporal constraint matrix \( T \) captures the temporal stability feature, which outlines that the change between two consecutive time slots is small. Hence, we set \( T = \text{Toeplitz}(0, 1, -1)_{t \times t} \). The Toeplitz matrix is defined with central diagonal given by 1, and the first upper diagonal given by -1, and the others given by 0.

This Toeplitz matrix adds the temporal constraint into the estimation. Importing \( \| L R^T T \|_F^2 \) into Eq. (14) is equal to induct an additional constraint into the original optimization problem. Since the temporal constraint is an inherent feature of environment, this additional constraint can filter more noises and errors in \( L R^T \) estimation.

**Spatial similarity improvement**: The spatial constraint matrix \( \mathbb{H} \) captures the spatial similarity feature, which reveals that values among one-hop neighbors nodes are usually similar. Hence, we set \( \mathbb{H} \) to be a row-normalized \( H^* \), where \( H^* = H + D \). The matrix \( H \) is a TM-1H, i.e., the one-hop topology matrix mentioned before. And \( D \) is an \( n \times n \) diagonal matrix, which is defined with central diagonal given by \( \text{diag}(d_1, d_2, \ldots, d_n) \), and the others given by 0. In \( D \), \( d_i = -\sum H(i) \).

The spatial similarity constraint is added by the matrix \( \mathbb{H} \). Computing the result of \( \mathbb{H} X \) is to get the differences between the elements and the average value of their one-hop neighbors in \( X \). As the same purpose of time improvement part, we introduce the part of minimizing \( \| \mathbb{H} L R^T \|_F^2 \) into Eq. (14). It takes advantage of the inherent environment feature as an additional constraint in optimization problem, which leads to a more accurate estimation of \( L R^T \), i.e., \( X \).

### 6.3 ESTI-CS Algorithm

We propose an efficient ESTI-CS algorithm to solve the estimation in the optimization problem Eq. (14).

First, we scale the \( T \) and \( \mathbb{H} \) as all \( \| \cdot \|_F^2 \) in Eq. (14) having the same order of magnitude. The scaling method is similar to [31]. Then ESTI-CS algorithm solves the optimization in an iterative manner. \( L \) is initialized randomly, so \( R \) can be computed by solving the following contradictory equation:

\[
\begin{bmatrix}
B \circ (L R^T) \\
\sqrt{\lambda} R^T
\end{bmatrix} =
\begin{bmatrix}
S \\
0
\end{bmatrix},
\tag{15}
\]

This equation can be rewritten as:

\[
\begin{bmatrix}
\text{Diag}(B(i)) L R^T_{(i)} \\
\sqrt{\lambda} R^T_{(i)}
\end{bmatrix} =
\begin{bmatrix}
S(i) \\
0
\end{bmatrix},
\tag{16}
\]

where \( i \) ranges from 1 to \( t \). This is a combination of multiple standard linear least squares problems. We then have \( R_{(i)} = (P^T_i P_i)^{-1}(P^T_i Q_i) \), where \( P_i = [\text{Diag}(B(i)) L; \sqrt{\lambda} R^T_{(i)}] \) and \( Q_i = [S(i); 0_r] \). Similarly, once \( R^T \) is obtained, \( L \) can be re-computed by fixing \( R^T \). This mutual re-computing process repeats until the optimal value is reached.

We analyze the complexity of the ESTI-CS algorithm. The key operation is the procedure for computing the inverse matrix, which provides the best approximate solution to the contradictory equation. This procedure is completed by a matrix multiplication [19]. Thus, its time complexity is \( O(r n t g) \). Since ESTI-CS repeats the procedure for \( g \) times, the total complexity is \( O(r n t g) \). From our evaluation experience in Sec. 8, \( L \) and \( R^T \) usually converge after 5 iterations.

### 7 Multi-Attribute Component

#### 7.1 MAA Overview

Multi-attribute assistant component can be utilized to improve the accuracy of ESTI-CS when correlation exists. Under such scenario, the proposed ERSN problem is extended to \( k \)-ERSN problem: Given \( K \) sensory matrices (SMs) \( S_k \), where \( k = 1, 2, \ldots, K \), and these SMs have the same size but different values. The goal is to jointly find the corresponding optimal reconstructed matrices (RMs) \( \hat{X}_k \) that approximate the original environmental matrices (EMs) \( X_k \).

For simplicity, we study the two-attribute situation as an example. Formally, when \( K = 2 \), ERSN problem is formulated as follows: Given \( S_1 \) and \( S_2 \), find an optimal solution for \( \hat{X}_1 \) and \( \hat{X}_2 \), i.e.,

\[
\begin{align*}
\text{Objective} & \quad \min (\| \hat{X}_1 - X_1 \|_F + \| \hat{X}_2 - X_2 \|_F), \\
\text{Subject to} & \quad B_1 \circ \hat{X}_1 = S_1, \\
& \quad B_2 \circ \hat{X}_2 = S_2. \\
\tag{17}
\end{align*}
\]

**Normalization**: Since the magnitudes of attributes are different, it may cause one matrix to overshadow another. In order to overcome this issue, \( X_1 \) and \( X_2 \) are normalized respectively.

**Low-Rank Matrix Approximation**: Eq.(17) is tied by \( X_1 \) and \( X_2 \), so the problem cannot be solved in closure form. However, due to the inherited low-rank feature, this problem can be converted to a rank minimization problem. Thus, the optimal \( \hat{X}_k \) is evaluated by the problem:

\[
\min (\text{rank}(\hat{X}_k)), \quad \text{s.t.} \quad S_k = \hat{X}_k \circ B_k. \\
\text{min (rank}(\hat{X}_k)), \quad \text{s.t.} \quad S_k = \hat{X}_k \circ B_k. \\
\tag{18}
\]

Still two problems are up against us: (1) the rank calculating operator \( \text{rank}(\cdot) \) is not convex. (2) there is no connection between \( X_1 \) and \( X_2 \).

To conquer the difficulty (1), we still utilize SVD-like factorization as \( \hat{X} = L R^T \). Thus \( \text{min (rank}(\hat{X})) \) is solvable by looking for \( L \) and \( R \), which satisfy \( \min (\|L\|_F^2 + \|R^T\|_F^2) \).
Compressive Sensing-based Joint Matrix Decomposition: To overcome the difficulty (2), we need to find the correlation between $X_1$ and $X_2$. Through the JSD analysis in Sec.5.3, we separate the approximation $X\hat{1}$ and $X\hat{2}$ by JSD as:

\[
\begin{align*}
X_1 &= \hat{W} + \Delta_1 \\
X_2 &= \hat{W} + \Delta_2
\end{align*}
\tag{19}
\]

Since $\hat{W}$, $\Delta_1$ and $\Delta_2$ inherit the low-rank feature, $k$-ERSN problem is reformulated as:

Objective \hspace{1cm} \min(||\hat{W}||_* + ||\Delta_1||_* + ||\Delta_2||_*),

Subject to \hspace{1cm} B_1 \circ (\hat{W} + \Delta_1) = B_1 \circ X_1, \tag{20}
B_2 \circ (\hat{W} + \Delta_2) = B_2 \circ X_2,

where $|| \cdot ||_*$ is the nuclear norm which is defined as the sum of singular values, e.g., $||X||_* = \sum_{i=1}^r \sigma_i(X)$.

Furthermore, using SVD-like factorization, $||\hat{W}||_* + ||\Delta_1||_* + ||\Delta_2||_*$ in Eq.(20) is rewritten as:

\[
||L_W||_F^2 + ||R_W^T||_F^2 + ||L_1||_F^2 + ||R_1^T||_F^2 + ||L_2||_F^2 + ||R_2^T||_F^2.
\tag{21}
\]

where $L_W$, $L_1$, $L_2$ are $n \times r$ matrices and $R_W$, $R_1$, $R_2$ are $r \times t$ matrices. Moreover, $\hat{W} = L_W R_W^T$, $\Delta_1 = L_1 R_1^T$, and $\Delta_2 = L_2 R_2^T$. For short, Eq.(21) is denoted by $\sum ||L_j||_F^2 + \sum ||R_j||_F^2$, where $j = 1, 2, W$.

To avoid overfitting, $k$-ERSN problem is rewritten to be a non-stationary optimization problem using the Lagrange multiplier method, i.e.,

\[
\begin{align*}
\min & \left( \lambda \sum ||L_j||_F^2 + \sum ||R_j||_F^2 \right) \\
+ & ||B_1 \circ (L_W R_W^T + L_1 R_1^T) - S_1||_F^2 \\
+ & ||B_2 \circ (L_W R_W^T + L_2 R_2^T) - S_2||_F^2. \tag{22}
\end{align*}
\]

Eq.(22) is the core of MAA component, which is solvable because (1) $B_1, B_2, S_1$ and $S_2$ are known, (2) each $|| \cdot ||_F^2$ is non-negative, (3) the optimal value can be reached by minimizing all non-negative parts to zero. Hence, $\hat{X}_1$ and $\hat{X}_2$ can be estimated by combining Eq.(22) and Eq.(19).

ESTI-CS with MAA is to reconstruct several (two as example) environments according to

\[
\begin{align*}
\min & \left( \lambda \sum ||L_j||_F^2 + \sum ||R_j||_F^2 \right) \\
+ & ||B_1 \circ (L_W R_W^T + L_1 R_1^T) - S_1||_F^2 \\
+ & ||B_2 \circ (L_W R_W^T + L_2 R_2^T) - S_2||_F^2 \\
+ & ||B_1 \circ (L_1 + L_W)(R_1 + R_W)^T - T_1||_F^2 \\
+ & ||B_2 \circ (L_2 + L_W)(R_2 + R_W)^T - T_2||_F^2. \tag{23}
\end{align*}
\]

It can be seen that Eq.(23) is the combination of Eq.(22) and Eq.(14). The first three items of Eq.(23) utilize the low-rank feature, which is the fundamental compressive sensing, the fourth and fifth items incorporate the spatial similarity improvement, the last two items merge the temporal stability improvement, and the multi-attribute assistant component is added a lastly into Eq.(23) by $L_j$ and $R_j$, where $j = 1, 2, W$.

Extension: The MAA component is also suitable for the case of more attributes. For instance, if we obtain $k$ attributes in one WSN, represented by $X_1, X_2, \cdots, X_k$. The utilization of MAA is to rewrite Eq.(20) into

\[
||\hat{W}||_* + ||\Delta_1||_* + ||\Delta_2||_* + \cdots + ||\Delta_k||_*. \tag{24}
\]

Then $k$-ERSN problem can be solved by the similar method of the above two-attribute case.

7.2 MAA Algorithm

We only present the core of MAA component in this section, which is to solve Eq.(22). And the realization of spatial-temporal improvement in Eq.(23) is the same method in ESTI-CS, we do not repeat it here.

The algorithm solves in an iterative manner. First, $L_1, L_2, L_W, R_1$, and $R_2$ matrices are initialized randomly. Then, $R_W$ can be calculated from the initialized matrices by solving the equation:

\[
\begin{bmatrix}
B_1 \circ (L_W R_W^T)
B_2 \circ (L_W R_W^T)
\end{bmatrix} = \begin{bmatrix}
S_1 - L_1 R_1^T
S_2 - L_2 R_2^T
\end{bmatrix}. \tag{25}
\]

Eq.(25) is solvable using the linear least square method. After $R_W$ is obtained, $L_W$ can be computed using the same procedure by fixing $R_W$. Similarity, any of $L_1, L_2, R_1$ and $R_2$ is computed by fixing the other three. Using the iterative manner, these four matrices can be obtained one-by-one. The computational complexity of ESTI-CS-MAA is the same as ESTI-CS.

8 PERFORMANCE EVALUATION

8.1 Methodology

The proposed ESTI-CS approach is compared with existing algorithms for missing data interpolation for environmental reconstruction in WSNs.

Ground truth. Since the performance evaluation needs complete EMs $X$ to compute the metric of error ratio (ER), we utilize the datasets as shown in Tab. 1. Six EMs are adopted: indoor-temp, indoor-light, forest-temp, forest-light, ocean-temp and ocean-light.

Methods. To verify the effectiveness of ESTI-CS, four classic interpolation methods are selected for comparison. They are compressive sensing (CS) [19] (computational complexity $O(rnt\varrho)$), Delaunay Triangulation (DT) [13] (computational complexity $O(nt \log nt)$), Multi-channel Singular Spectrum Analysis (MSSA) [32] (computational complexity $O(rnt\log nt + r^2nt)$), and K-Nearest Neighbor (KNN) [7] (computational complexity $O(nt)$). The parameter $K$ in KNN is set to be $\sum_{i=1}^n H_{(i)}/n$. The parameter $M$ in MSSA is set to 32 as suggested by [32].

Procedure. The procedure of simulation is: 1) Generate BIM $B$ according to four loss patterns. 2) Compute SM $S$ according to Definition 3: $S = B \circ X$. 3) All interpolation algorithms being tested take SMs as
input and generate RMs. 4) The accuracy metric ER is computed between EMs and RMs. And finally, these errors are compared.

Series. Three series of experiments are evaluated. The basic experiment measures the performance of different algorithms against typical random loss probability. The second experiment evaluates the performance in diverse data loss patterns. And the third experiment compares the performance of ESTI-CS with MAA and ESTI-CS without MAA.

8.2 Comparison on Random Loss Pattern
In the basic comparison, we test the error ratios under diverse algorithms on the element random loss (ERL) pattern only. The data loss rate \( p_{ERL} \) ranges from 10% to 90%. Fig. 3 shows the results. The X-axis presents the data loss probability, and the Y-axis is the value of ER, which represents the reconstruction accuracy.

In the indoor-temp Fig. 3(a), ESTI-CS shows the best performance. Even 90% data have been lost, ESTI-CS still can reconstruct the environment with \( ER \leq 10\% \). While ER of CS is about 19%, DT is close to 38%, and ER of KNN and MSSA are more than 60%. ESTI-CS is much better than other algorithms in this scenario. In the indoor-light Fig. 3(b), ESTI-CS still outperforms the others, but the advantage is less significant than that in indoor-temp. The reason is that the indoor temperature change has strong spatio-temporal feature. However, the change of indoor light is largely influenced by the light switch. So the indoor light dataset observes more artificial changes than spatio-temporal stability.

The performance of Forest-Temp and the Forest-Light are similar. The reason is that both the temperature and the light are mainly affected by the sun due to an outdoor application. These two environment attributes have strong correlation. As shown in Fig. 3(c) and Fig. 3(d), ESTI-CS achieves the best environment reconstruction among the five algorithms. CS, MSSA and DT fall behind ESTI-CS a little. KNN is not bad...
when $p_{ERL} < 50\%$, but when $p_{ERL} > 50\%$, ER of KNN increases quickly.

In the ocean-temp Fig. 3(e), ESTI-CS and DT produce the similar performance. When the data loss is 90\%, they achieve ER < 30\%. Meanwhile, the ERs of CS, KNN and MSSA are bigger. In the ocean-light Fig. 3(f), the performance of ESTI-CS and DT are similar with the range of loss rate from 10\% to 80\%. When the loss rate increases to 90\%, ER of DT also increases rapidly, and ER of ESTI-CS still keeps within 22\%. These two figures indicate that ESTI-CS perform better than DT, CS, KNN and MSSA in this outdoor and small-scale WSN scenario.

Overall, ESTI-CS obtain lower interpolation error, which can be used in almost all tested datasets with different loss ratios. KNN and DT produce similar but the poor ER performance, because both of them interpolate with only the space relation among nodes but no time relation consideration. CS and MSSA are better than KNN and DT, but still worse than ESTI-CS. Especially, at the high data loss cases (data loss $\geq 80\%$), ESTI-CS exhibits an evident advantage over other algorithms. In all dataset, ESTI-CS can successfully achieve an environment reconstruction with 20\% error when there are 90\% data are missing.

8.3 Performance on Data Loss Patterns

In Fig. 4, we plot the comparison histograms of five algorithms for reconstructing the environment with different data loss patterns.

In the simulation for BRL pattern, each of the six EMs is set to lose data with the block pattern. The scale and the number of the blocks are random, but the amount of total data loss is 50\% in this simulation. In Fig. 4(a), most algorithms in most EMs perform not well. For example, in forest-light, ER of all algorithms are bigger than 60\%. The reasons are 1) in the forest, many shadows disturb the spatio-temporal stability. 2) if large blocks of data lose, spatio-temporal optimized estimation is helpless either. These two reasons lead to the result. However, in indoor-temp, ocean-temp, and ocean-light, the environment changes are smoothly, ER of ESTI-CS are less than 5\% despite 50\% BRL data loss. Even indoor-light, forest environments, ESTI-CS is still a bit better than the others. In addition, we find that KNN is in big trouble for estimating the missing data in BRL.

In the simulation of EFLR pattern, the rows are randomly selected, the loss frequency in these rows is set \( > 75\% \), and the totally lose data in matrix is set 50\%. We find that the results in Fig. 4(b) are close to the basic comparison, because the data loss in EFLR is similar to ERL pattern. In EFLR, the temporal optimization can contribute a partial effect, but the space optimization still works. So our ESTI-CS still outperforms CS, KNN, DT, and MSSA.

The performance of SELR is similar to EFLR, we show the SELR result in the supplemental document.

In the simulation of Combinational Loss pattern, we set 20\%ERL + 10\%BRL + 10\%EFLR + 10\%SELR. The results of five algorithms are shown in Fig. 4(c). The ER of ESTI-CS is $\leq 20\%$ in any dataset in the combinational loss pattern.

In summary, ESTI-CS outperforms CS, KNN, DT and MSSA in any data loss pattern.

8.4 Performance on ESTI-CS with MAA

In this series, we evaluate the benefit from MAA component for ESTI-CS. The data loss pattern and loss rate setting are the same as the setting in Sec. 8.2.

In Fig. 5, we illustrate the ER results of ESTI-CS-MAA algorithm and compare them with ESTI-CS in the case of two attributes under random loss pattern. In every dataset, two attributes temperature and light are reconstructed together by ESTI-CS-MAA. It can be seen that the reconstruction accuracy of ESTI-CS-MAA is universally better than that of ESTI-CS in three real WSN datasets.
Fig. 5(a) and Fig. 5(b) show ESTI-CS-MAA is slightly better than ESTI-CS in Intel Indoor dataset. In face of 90% data loss, ESTI-CS-MAA improves the ER by 2% in Indoor-Temp shown in Fig. 5(a) and 3% shown in Indoor-Light in Fig. 5(b) compared with ESTI-CS. In GreenOrbs dataset, ESTI-CS-MAA performs better than ESTI-CS. The MAA-enabled algorithm outperforms baseline ESTI-CS by 7% in Forest-Temp shown in Fig. 5(c) and 6% in Forest-Light shown in Fig. 5(d).

The results of ESTI-CS-MAA are significantly better than those of ESTI-CS in OceanSense dataset. As shown in Fig. 5(e) and Fig. 5(f), 14% and 12% ER are improved by ESTI-CS-MAA respectively in Ocean-Temp and in Ocean-Light.

Recall the correlation analysis of Fig. 2, the correlation between temperature and light in OceanSense is higher than that in GreenOrbs and even higher than that in Intel Indoor. We summarize that MAA can further improve the ESTI-CS on reconstruction accuracy when multi-attribute correlation exists. Moreover, the higher the correlation is, the more improvement the MAA provides.

9 Conclusion
In this paper, we studied the data loss and reconstruction problem in WSNs. We verified the massive data loss in real datasets and modeled the special data loss patterns of WSNs. Then, we mined the low-rank, spatial, temporal, and correlation features from WSN datasets. By drawing on these observations, we designed ESTI-CS with MAA algorithm to estimate the missing data. The proposed algorithm combines the benefits of compressive sensing, environmental space-time, and multi-attribute correlation features. Trace-driven experiments illustrated that ESTI-CS outperforms existing interpolation methods.

Acknowledgment
This research was supported by NSFC under grant No. 61303202, No. 61073158, No. 61102010, STCSM Project No. 12dz1507400, Doctoral Program Foundation under grant No. 201107132012001, FQRTN grant 131844, Singapore-MIT IDC IDD61000102a, SUTD-ZJU/RES/03/2011, and NRF2012EW7-ERIP002-045.

References
Linghe Kong (S’09-M’13) is currently a research assistant professor in Shanghai Jiao Tong University and a postdoctoral researcher in Singapore University of Technology and Design. He received his PhD degree in Computer Science with Shanghai Jiao Tong University 2012, Dipl. Ing. degree in Telecommunication with TELECOM Sud-Paris 2007, B. E. degree in Automation with Xidian University 2005. His research interests include wireless sensor networks and mobile computing.

Mingyuan Xia received the B.E. degree in Computer Science and Engineer from Shanghai Jiao Tong University, China, in 2011, and is now a PhD candidate in the School of Computer Science at McGill University, Canada. His research interests include wireless sensor networks and operations systems.

Xiao-Yang Liu received his B. E. degree in computer science in Huazhong University of Science and Technology, China, in 2010. He is currently working toward the PhD degree in Department of Computer Science and Engineer in Shanghai Jiao Tong University. His research interests are in the area of wireless sensor networks.

Guangshuo Chen received the B.E. degree in Electronic and Information Engineering from Shanghai Jiao Tong University, China, in 2011, and is now an academic Master degree candidate in Computer Science from Shanghai Jiao Tong University. His research interests include Wireless Sensor Networks (WSNs), Body Area Networks and Vehicular Networks. At present, he focus on the data recovery problem in WSNs.

Min-You Wu (S’84-M’85-SM’96) received the M.S. degree from the Graduate School of Academia Sinica, Beijing, China, in 1981 and the Ph.D. degree from Santa Clara University, Santa Clara, CA in 1984. He is a professor in the Department of Computer Science and Engineering at Shanghai Jiao Tong University and a research professor with the University of New Mexico. He serves as the Chief Scientist at Grid Center of Shanghai Jiao Tong University. His research interests include grid computing, wireless networks, sensor networks, multimedia networking, parallel and distributed systems, and compilers for parallel computers.

Xue Liu received the B.S. degree in mathematics and the MS degree in automatic control both from Tsinghua University, China, and the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2006. He is an associate professor in the School of Computer Science at McGill University. He has also worked as the Samuel R. Thompson associate professor in the University of Nebraska-Lincoln and HP Labs in Palo Alto, California. His research interests include computer networks and communications, smart grid, real-time and embedded systems, cyber-physical systems, data centers, and software reliability.

Supplemental Document: Data Loss and Reconstruction in Wireless Sensor Networks

Linghe Kong, Member, IEEE, Mingyuan Xia, Xiao-Yang Liu, Guangshuo Chen, Yu Gu, IEEE Member, Min-You Wu, IEEE Senior Member, Xue Liu, IEEE Member

Abstract—Reconstructing the environment by sensory data is a fundamental operation for understanding the physical world in depth. A lot of basic scientific work (e.g., nature discovery, organic evolution) heavily relies on the accuracy of environment reconstruction. However, data loss in wireless sensor networks is common and has its special patterns due to noise, collision, unreliable link, and unexpected damage, which greatly reduces the reconstruction accuracy. Existing interpolation methods do not consider these patterns and thus fail to provide a satisfactory accuracy when missing data become large. To address this problem, this paper proposes a novel approach based on compressive sensing to reconstruct the massive missing data. Firstly, we analyze the real sensory data from Intel Indoor, GreenOrbs, and Ocean Sense projects. They all exhibit the features of low-rank structure, spatial similarity, temporal stability and multi-attribute correlation. Motivated by these observations, we then develop an environmental space time improved compressive sensing (ESTI-CS) algorithm with a multi-attribute assistant (MAA) component to optimize the missing data estimation. Finally, the extensive simulation results on real sensory data sets show that the proposed approach significantly outperforms existing solutions in terms of reconstruction accuracy.

Index Terms—Wireless Sensor Networks, Data Loss and Reconstruction, Compressive Sensing

1 INTRODUCTION

People investigate the environment in order to understand our physical world. Recently, wireless sensor networks (WSNs) [15], [20] are widely adopted to gather sensory data and reconstruct the environment in the cyber space [11]. There are plenty of environment monitoring applications under the water [30], in the forest [23], and on the volcano [28]. An environment matrix (EM) is a common way to represent a dynamic environment. An EM is an $n \times t$ matrix that records data from $n$ sensors over $t$ time intervals. Environment reconstruction [13] attempts to obtain the full and accurate EM from raw sensory data, which is an essential step for any further analysis.

Motivation: A great deal of basic scientific work heavily depends on the accuracy of environment reconstruction. For example, scientists reveal the nature of ocean currents from accurate underwater temperature data [30], understand the demand for plant evolution based on the light condition in the forest [23], and discover the eruption omen by monitoring the shake of the volcano [28].

However, since data gathering is largely affected by hardware and wireless conditions, a raw dataset usually has notable missing data. Furthermore, missing data become larger as WSNs grow in scale [3]. Consequently, data loss in WSNs becomes the key challenge against accurate environment reconstruction. It is urgent and important to design effective methods to recover incomplete EMs.

Existing approaches and limitations: The missing value problem is fundamental in dataset field. Lots of work has been contributed such as the local interpolation method K-Nearest Neighbors (KNN) [7], the global refinement method Delaunay Triangulation (DT) [13], and the principal component analysis method Multi-channel Singular Spectrum Analysis (MSSA) [32]. These methods are often used when there are only a few missing values, but cannot be applied when the missing data grow.

Compressive Sensing (CS) [5], [8] is a powerful and generic technique for estimating missing data. CS can recover an entire dataset from only a small fraction of data as long as these data contain sparse/low-rank features. So far, CS has been utilized to reconstruct network traffic [31], refine localization [24] and improve urban traffic sensing [19]. However, since a WSN has unique data loss patterns, directly applying CS on EM interpolation cannot gain satisfactory accuracy.

Our work and contribution: Our work is fourfold:

Firstly, we analyze real environmental data from Intel Indoor [1], GreenOrbs [23], and OceanSense [30] projects. We confirm the massive data loss in general applications and mine the specific data loss patterns...
in WSNs. Then we reveal four features in environmental datasets: 1) **Low-rank structure.** A complete EM can be represented by a few principle data, which underpins the applicability of CS. 2) **Time stability.** The sensory values of one certain node are usually similar at adjacent time slots. 3) **Space similarity.** The sensory values of neighbor nodes are similar. 4) **Multi-attribute correlation.** Multiple environmental attributes have strong correlation in some cases. For example, the change tendency of temperature and light are have strong correlation in some cases. For example, the change tendency of temperature and light are

Multiple environment reconstruction of WSN data is required to discover four data loss patterns. We model the environment with EM and reveal four data loss patterns. Typically, ESTI-CS can achieve an effective environment reconstruction with less than 20% error when there are 90% missing data in the collected data. In addition, MAA further enhances the performance of ESTI-CS in extensive simulation.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first work to study the data loss and reconstruction in WSNs. We model the environment with EM and discover four data loss patterns.
- We mine several large WSN datasets and reveal the features of low-rank structure, time stability, space similarity, and multi-attribute correlation.
- Based on the observed features, we design the ESTI-CS algorithm to accurately estimate the missing data in highly incomplete EM.
- We design an optional component MAA algorithm to further improve the accuracy of ESTI-CS when applicable.
- The proposed ESTI-CS is simulated based on real data. The performance shows the ESTI-CS w/o MAA is effective for massively data loss against existing interpolation methods.

**Paper organization:** Section 2 presents the related work. Section 3 models the problem. Section 4 analyzes the data loss. Section 5 mines the environmental features. Section 6 proposes the ESTI-CS algorithm. Section 7 presents the optional component MAA. Section 8 evaluates the proposed approach. Section 9 concludes the paper.

## 2 Related Work

The missing value problem is common in datasets [3]. A great deal of existing work has devoted to interpolating the missing data. K-Nearest-Neighbor (KNN) [7] is a classical local interpolation method. KNN simply utilizes the values of the nearest K neighbors to estimate the missing one. It is frequently used in many low-fidelity estimation cases. Delaunay Triangulation (DT) [13] is a typical global refinement method, which treats the gathered data as vertices. DT takes advantage of these vertices and their global errors to build virtual triangles for data interpolation. It is widely adopted in computer vision for surface rendering. Multi-channel Singular Spectrum Analysis (MSSA) [32] is a data adaptive and nonparametric method based on the embedded lag-covariance matrix. MSSA is often used in geographic data recovery.

Despite much progress in the area of data interpolation, existing methods are suitable for only few missing values, but perform poorly when the loss rate grows high, which is common in WSN cases.

Compressive sensing (CS) is an advanced method to recover the whole condition with just a few sampled data [6], [8]. Its fundamental theory has been utilized in plenty of fields such as statistics, image processing, signal recovery, and machine learning. As for missing value estimation, CS-based methods have been developed for network traffic estimation [31], road traffic interpolation [19], and localization in mobile networks [24]. CS has also witnessed wide applications in WSNs, e.g., recovering signal under noisy background [2], balancing load via compressive data gathering [22]. However, the study of CS for environment reconstruction in WSNs is still vacant.

Existing CS-based interpolation methods cannot be directly applied for accurate environment reconstruction due to two reasons: 1) CS-based methods require the dataset to have inherent structure. Features that are extracted from network traces [31] or road traffic [19] are not applicable for WSN data. 2) CS theory performs well when the missing values follow the Gaussian or pure random distribution [18], [27]. However, as shown in Section 4.3, the loss patterns of WSNs do not satisfy these prerequisites.

To address the above challenges, an effective environment reconstruction of WSN data is required to consider massive data loss as well as to study WSN-specific loss patterns.

## 3 Problem Formulation

### 3.1 Environmental Data Reconstruction

Rebuilding the virtual environment (such as the dynamic temperature, light, gas concentration, or magnetic strength in real world) in cyber space based on the sensory data is called environment reconstruction.

In environment reconstruction systems, sensor nodes are scattered in the given area. Suppose totally
Problem Statement

Data reconstruction is to rebuild the real environment (EM) based on the gathered sensory data (SM).

Definition 4 Reconstructed Matrix (RM): is generated by interpolating the missing values in an SM to approximate EM. RM is denoted by $\hat{X} = (\hat{x}(i,j))_{n \times t}$.

**Problem: Environment Reconstruction in Sensor Network (ERSN):** Given an SM $S$, the ERSN problem is to find an optimal RM $\hat{X}$ that approximates the original EM $X$ as closely as possible. i.e.,

**Objective:** $\min_S \| X - \hat{X} \|_F$, Subject to: $S$,

where $\| \cdot \|_F$ is the Frobenius norm used to measure the error between matrix $X$ and $\hat{X}$. For calculating, take $X$ as an example, $\| X \|_F = \sqrt{\sum_{i,j} (x(i,j))^2}$.

In the ERSN problem, the objective is to minimize the absolute error. In order to measure the reconstruction error in different scenarios among different methods, we further define the following metric.

Definition 5 Error Ratio (ER): is the metric for measuring the reconstruction error after interpolation:

$$\epsilon = \frac{\sqrt{\sum_{i,j;b(i,j)=0} (x(i,j) - \hat{x}(i,j))^2}}{\sqrt{\sum_{i,j;b(i,j)=0} (x(i,j))^2}}.$$  
(4)

Note that the condition $b(i,j) = 0$ in Eq. (4) indicates that only errors on the missing data are counted.

1. In this paper, $AB$ presents the matrix production of $A$ and $B$. $A \circ B$ presents the element-wise production of $A$ and $B$.

### 4 Data Loss in Sensor Networks

#### 4.1 Environmental Datasets

In this section, we analyze the data loss in real WSN datasets. The three datasets are Intel indoor, GreenOrbs, and OceanSense projects.

The data of Intel indoor experiment [1] are gathered by Intel Berkeley Research lab from February 28th to April 5th, 2004. There are 54 Mica2Dot nodes placed in a $40\text{m} \times 30\text{m}$ room. Every node reports once every 30 seconds. Sensory data include temperature, light, and humidity.

GreenOrbs project [23] deploys a real WSN for forest surveillance from 2008 to present. More than 450 TelosB nodes are scattered on the Tianmu Mountain, China and gather temperature, light, and humidity once every 10 minutes.

Ocean Sense project [30] carried out in the sea of Taipingjiao, China from 2007 to present. This dataset contains 20 TelosB nodes monitoring an area of $300\text{m} \times 100\text{m}$. Each sensing node reports temperature and light data every 2 minutes.

#### 4.2 Massive Data Loss

Through statistics analysis, we verify that the significant data loss exists in all of these original datasets.

We investigate totally 54 nodes and 84,600 time slots (one month) data from the Intel Indoor dataset. 23% data points are missing. The GreenOrbs dataset also observes 35% data loss. And this loss is even larger in OceanSense, which is about 64% for 20 nodes and 5,040 time slots (one week). The basic information of three datasets and their data loss ratios are listed in Table 1. We find that the data loss is common and significant in real WSNs.

#### 4.3 Data Loss Pattern

Traditional work usually assumes that the data loss follows a random distribution [19], [32]. However, this claim does not apply to the WSN situation. In terms of the nature of WSNs, we synthesize several typical data loss patterns.

**Pattern 1 Element Random Loss (ERL):** This is the simplest loss pattern. Data elements in the matrix are dropped independently and randomly. As shown in Fig. 1(a), missing data points for ERL are randomly distributed in a SM. The noise and collision [12] in WSNs are the root causes of random element loss.

**Pattern 2 Block Random Loss (BRL):** Data from adjacent nodes in adjacent time slots (form a block) are
dropped together. In WSNs, congestion [10] always leads to data loss on high-density sensor nodes during a period of time. Fig. 1(b) visualizes this scenario.

**Pattern 3 Element Frequent Loss in Row (EFLR):** Unreliable links [29] are common phenomenon in real wireless scenarios. When the link quality is not good, sensory data are prone to loss due to the intermittent transmission. As shown in Fig. 1(c), in EFLR, elements in some particular rows have a higher missing probability.

**Pattern 4 Successive Elements Loss in Row (SELR):** This pattern models that a given node starts losing from a particular time slot. This type of loss occurs when some sensor nodes are damaged or run out of energy [26], which is made visible by Fig. 1(d).

**Pattern 5 Combinational Loss (CL):** In real world, data loss is a combination of loss patterns above.

### 5 Environmental Data Mining

#### 5.1 Ground Truth

In order to discover the environmental features, the complete datasets are desired as the ground truth. However, EMs from the three original datasets cannot be directly utilized since they all have considerable data loss. To generate applicable EMs, we perform preprocessing on the raw datasets as demonstrated in Fig. 2, which selects the small but complete subsets from these three datasets. The size and time interval of selected matrices are shown in Table 2. As a result, six EMs are generated from preprocessing: indoor temperature, indoor light, forest temperature, forest light, ocean temperature, and ocean light.

#### 5.2 Low-Rank Structure Discovery

Environmental data of different locations over different times are not independent. There exists inherent structure or redundancy. We mine these features in above selected datasets by Principal Component Analysis (PCA), which is an effective non-parametric technique for revealing the hidden structure that often underlies a dataset [16].

Any \( n \times t \) matrix \( X \) can be decomposed into three matrices according to Singular Value Decomposition (SVD) method:

\[
X = U \Sigma V^T = \sum_{i=1}^{\min(n,t)} \sigma_i u_i v_i^T, \tag{5}
\]

where \( V^T \) is the transpose of \( V \), \( U \) is an \( n \times n \) unitary matrix (i.e., \( U U^T = U^T U = I_{n \times n} \)), \( V \) is a \( t \times t \) unitary matrix (i.e., \( V V^T = V^T V = I_{t \times t} \)), and \( \Sigma \) is an \( n \times t \) diagonal matrix constraining the singular values \( \sigma_i \) of \( X \). Typically, the singular values in \( \Sigma \) are sorted, i.e., \( \sigma_i \geq \sigma_{i+1} \), \( i = 1, 2, \ldots, \min(n,t) \), where \( \min(n,t) \) is the number of singular values. The rank of a matrix, denoted by \( r \), is equal to the number of its non-zero singular values. If \( r < \min(n,t) \), the matrix is low-rank.

In Eq. (5), the singular value \( \sigma_i \) also indicates the energy of the \( i \)-th principal component. The total energy is equal to the sum of all singular value \( \sum_{i=1}^{\min(n,t)} \sigma_i \). According to PCA, a low-rank matrix [31] exhibits that its first \( r \) singular values occupy the total or near-total energy \( \sum_{i=1}^r \sigma_i \approx \sum_{i=1}^{\min(n,t)} \sigma_i \).

In Fig. 3(a), we illustrate the distribution of singular values in 6 EMs. The X-axis presents the \( t \)-th singular values. Since the scales of 6 EMs are different, we normalize the X-axis. So \( \min(n,t) \) of every EM is normalized to 100%. The Y-axis presents the values of the sum of first \( i \)-th singular value. Due to the same reason of X-axis, the Y-axis is also normalized. i.e., \( \max(\sigma_i) \) of every EM is normalized to 1. This figure suggests that the energy is always contributed by the top several singular values in real environments. For example, the top 5% singular values contribute all energy in Indoor-Temp; the top 12% \( \sigma_i \) include...
all energy in Forest-Temp; and even in the worst case of Ocean-Light, the top 25% singular values contribute the most of energy. The universal existence of $\sum_{i=1}^{n} \sigma_i \approx \sum_{i=n+1}^{m} \sigma_i$, and $r \ll \min(n,t)$ reveals that EMs exhibit obvious low-rank structures. Low-rank features [19] serve for the prerequisite for using compressive sensing.

5.3 Temporal Stability Feature

In real world, most of sensory data (e.g., temperature) change stably, i.e., there is little mutation on environmental value between adjacent time slots. On the basis of this natural phenomenon, we analyze the datasets in time dimension to reveal temporal features.

We measure the temporal stability at node $i$ and time slot $j$ by computing the normalized difference values between adjacent time slots $\Delta x(i,j)$:

$$\Delta x(i,j) = \frac{|x(i,j) - x(i,j-1)|}{\max(|x(I,J) - x(I,J-1)|)}$$

(6)

where $I$ varies from 1 to $n$, $J$ varies from 1 to $t$, and $\max(|x(I,J) - x(I,J-1)|)$ is the maximal difference between any two consecutive time slots in the EM.

The CDF of $\Delta x(i,j)$ is plotted in Fig. 3(b). The X-axis presents the normalized difference values between two consecutive time slots, i.e., $\Delta x(i,j)$. The Y-axis presents the cumulative probability. We observe that > 80% in the Forest datasets, > 60% in the Indoor datasets, and > 50% in the Ocean-Temp, the value of $\Delta x(i,j)$ is 0. i.e., the environmental value is not changed between two consecutive time slots. In addition, near all (> 95%) $\Delta x(i,j)$ are very small (< 0.05) in Forest and Indoor datasets. Even in the worst case, the ocean-light values between two consecutive time slots mostly (> 80%) change only a little (< 0.3). These results indicate that temporal stability exists in real environments. Based on this discovery, we can adopt the time feature to optimize the compressive sensing technique for missing data estimation.

5.4 Spatial Similarity Feature

We also consider the difference value from the space dimension. We know that environments are often smooth in a small area, i.e., at the same time, environmental values are similar at nearby locations.

In real WSN applications, the locations of nodes can either be known [23] or unknown [30]. Generally, it is not easy to know the actual distance between nodes from a WSN without GPS information. Although physical distance may not be available, the network topology is always easy to obtain. The topology can be known from the routing information when the sink gathers sensory data from nodes. Constrained by the wireless power, sensors are usually located near their one-hop neighbors. First, Topology Matrix for One-Hop (TM-1H) nodes $H$ is defined as:

$$H = (h(y,z))_{n \times n} = \begin{cases} 1 & \text{if } y \text{ and } z \text{ are neighbors;} \\ 0 & \text{otherwise,} \end{cases}$$

(7)

where $y = 1, 2, \ldots, n$, $z = 1, 2, \ldots, n$. Both rows and columns in a TM-1H represent sensor nodes, and $h(y,z)$ represents whether the node $y$ and node $z$ are one-hop neighbor or not. The TM-1H demonstrates the binary relationship between nodes.

Then, the spatial similarity at node $i$ and time slot $j$ is measured by computing the normalized difference between the value of a node and the average value of its all one-hop neighbors $\nabla x(i,j)$:

$$\nabla x(i,j) = \frac{x(i,j) - (H(i)X^{(i)}/\sum H(i))}{\max(x(I,J)) - \min(x(I,J))}$$

(8)

where $H(i)$ is the $i$-th row of matrix $H$, $X^{(i)}$ is the $j$-th column of matrix $X$. $H(i)X^{(i)}$ depicts the sum values of all one-hop neighbors of node $i$ at time slot $j$. $\sum H(i)$ represents the number of one-hop neighbors of node $i$. $\max(x(I,J))$ and $\min(x(I,J))$ are the maximum and minimum environmental value in the EM, and $\max(x(I,J)) - \min(x(I,J))$ stands for the maximal difference value.

The CDF of $\nabla x(i,j)$ is plotted in Fig. 3(c). The X-axis presents the normalized difference between value of one node and the average value of its all one-hop neighbors, i.e., $\nabla x(i,j)$. The Y-axis presents the cumulative probability. We find that no matter in which dataset, > 95% the value of $\nabla x(i,j)$ is < 0.3.
These results imply that real environments also have the feature of spatial similarity, i.e., the value of a node is similar to the value of its neighbors. Thus, the compressive sensing based estimation approach can be also optimized by space feature.

5.5 Multi-Attribute Correlation

We are aware of the following two facts in real WSN applications. (1) Usually, WSNs gather multiple attributes simultaneously, e.g., a TelosB node [21] senses three environmental attributes: temperature, light and humidity. (2) Multiple attributes have correlations in some applications. For instance, in the outdoor application, the attributes of temperature and light increase together when the sun is arising. In addition, the empirical study [17] reveals that several attributes do have relationship such as relative humidity and dewpoint temperature. Hereby, we propose to mine and exploit such correlations to further optimize environment reconstruction.

**Joint Sparse Decomposition:** In order to mine the correlations, a Joint Sparse Decomposition (JSD) method is proposed to jointly divide multi-attribute EMs into a public sub-matrix W and multiple private sub-matrices Δ. All sub-matrices have the same size of EMs, but their magnitudes are smaller. Suppose two attributes \( X_1 = (x_1^{(1)}, \ldots, x_1^{(t)}) \) and \( X_2 = (x_2^{(1)}, \ldots, x_2^{(t)}) \), where \( x_k^{(j)} \) presents the \( j \)-th column vector of EM \( X_k \), \( j = 1, 2, \ldots, t \). For both column vector \( x_1^{(j)} \) and \( x_2^{(j)} \), the goal is to split them into:

\[
\begin{align*}
    x_1^{(j)} &= w^{(j)} + \delta_1^{(j)}, \\
    x_2^{(j)} &= w^{(j)} + \delta_2^{(j)}, \\
    w^{(j)} &= \Psi v^{(j)},
\end{align*}
\]

where \( w^{(j)} \) is the public sub-vector of \( x_1^{(j)} \) and \( x_2^{(j)} \), which is the multiplications of a wavelet basis \( \Psi \) [4] and a sparse vector \( v^{(j)} \) satisfying \( w^{(j)} = \Psi v^{(j)} \). The private sub-vectors are represented by \( \delta_1^{(j)} \) and \( \delta_2^{(j)} \) respectively. According to Compressive Sensing theory [8][5], \( v^{(j)}, \delta_1^{(j)}, \) and \( \delta_2^{(j)} \) are obtained by solving an \( l_1 \)-norm minimization problem as the following:

\[
\hat{\theta} = \arg\min ||\theta||_1, \quad s.t. \quad x = A\theta.
\]

![Fig. 4. Correlation analysis by JSD. High energy fraction of W indicates strong correlation.](image)

**TABLE 3**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sub-Matrix</th>
<th>Inherited Low-rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Indoor</td>
<td>W</td>
<td>7%</td>
</tr>
<tr>
<td>-temp</td>
<td>( \Delta_1 )</td>
<td>2%</td>
</tr>
<tr>
<td>-light</td>
<td>( \Delta_2 )</td>
<td>8%</td>
</tr>
<tr>
<td>GreenOrbs</td>
<td>W</td>
<td>9%</td>
</tr>
<tr>
<td>-temp</td>
<td>( \Delta_1 )</td>
<td>9%</td>
</tr>
<tr>
<td>-light</td>
<td>( \Delta_2 )</td>
<td>12%</td>
</tr>
<tr>
<td>OceanSense</td>
<td>W</td>
<td>17%</td>
</tr>
<tr>
<td>-temp</td>
<td>( \Delta_1 )</td>
<td>11%</td>
</tr>
<tr>
<td>-light</td>
<td>( \Delta_2 )</td>
<td>18%</td>
</tr>
</tbody>
</table>

where \( || \cdot ||_1 \) is the \( l_1 \)-norm, \( \vartheta = (\psi^{(j)T}, \delta_1^{(j)T}, \delta_2^{(j)T})^T \), \( x = (x_1^{(j)T}, x_2^{(j)T})^T \) and \( A = (\Psi, I, 0; \Psi, 0, I) \). It was proved in [8] that solving Eq.(10) is NP-hard, so we adopt the least angle regression method, which was proposed in [9], to obtain \( \vartheta \). Then the public sub-vector \( w^{(j)} \), the private sub-vectors \( \delta_1^{(j)} \), and \( \delta_2^{(j)} \) can be calculated from \( \vartheta \).

Applying JSD onto every column vector, \( X_1 \) and \( X_2 \) are decomposed as

\[
\begin{align*}
    X_1 &= W + \Delta_1, \\
    X_2 &= W + \Delta_2.
\end{align*}
\]

**Correlation:** The energy fraction of public sub-matrix \( W \) is used to measure the correlation between two attributes, where the total energy is the sum of all singular values of three sub-matrices \( W, \Delta_1, \) and \( \Delta_2 \). Fig.4 shows the energy fraction of sub-matrices after JSD in diverse groups. Group (a) shows the results of JSD on two irrelevant random matrices. The public sub-matrix \( W \) contains only 7% of total energy. Group (b) shows the results of indoor-temp and indoor-light. The change of indoor-light has mutations by manually switching lights on/off, which leads to low correlation with indoor-temp \( W = 11\% \). The results of forest-temp and forest-light are shown in group (c). Both outdoor light and temperature vary according to the sun. However, due to the influence of tree shade, the correlation is not very strong. So \( W \) contains 29% of total energy, while the private sub-matrices \( \Delta_1 \) and \( \Delta_2 \) contain 35% and 36% respectively. And sensor nodes are fully exposed under the sun in OceanSense. Hence, high correlation between ocean-temp and ocean-light is shown in group (d), where \( W = 46\% \). Two same matrices have definitely highest correlation. When JSD is operated on two same EMs \( X_1 = X_2, W \) contains 100% energy and \( \Delta_1 = \Delta_2 = 0 \) in group (e). Fig.4 validates that JSD can be utilized to measure the correlation between two matrices. In addition, the higher correlation between two EMs, the more energy fraction contains in the public sub-matrix.

**Inherited Low-rank:** decomposed from EMs by applying JSD, the sub-matrices \( W, \Delta_1 \) and \( \Delta_2 \) still exhibit the low-rank features.

All sub-matrices are decomposed respectively by
SVD method. The same method mentioned in Sec.5.2 is adopted to determine the inherited low-rank feature. In Tab.3, we show the percentage of singular values that contain 90% of the total energy. As shown, for all the dataset, 7% to 20% of the top singular values can concentrate 90% of the total energy. Hence, the inherited low-rank features are exhibited in sub-matrices, which indicates that any of \( W, \Delta_1 \) and \( \Delta_2 \) can be recovered by CS based method.

Correlation and inherited low-rank motivate us to improve ESTI-CS by multi-attribute correlation.

6 Environmental Space Time Improved Compressing Sensing Approach

We propose a novel missing data estimation approach to address the ERSN problem. The proposed algorithm, namely environmental space time improved compressive sensing (ESTI-CS), takes into consideration the spatio-temporal features to optimize the estimation accuracy.

6.1 Compressive Sensing based Approach

Compressive sensing, which can tolerate high data loss, is a potential approach for the proposed ERSN problem. Mathematically, CS based approach can only be applied to sparse matrices. Furthermore, a low-rank matrix can be well approximated by a sparse matrix. Since we have revealed the low-rank structure in most real environment datasets, we propose to use CS method to estimate missing data from the SM.

The goal of solving the ERSN problem is to estimate \( \hat{X} \). According to Eq. (5), any matrix can be decomposed by SVD into \( \sum_{i=1}^{\min(n,t)} \sigma_i u_i v_i^T \). Through the inverse process, we can also create a r-rank approximation \( \hat{X} \) by using only the \( r \) largest singular values and abandoning the others:

\[
\sum_{i=1}^{r} \sigma_i u_i v_i^T = \hat{X}.
\]

This \( \hat{X} \) is known as the best r-rank approximation that minimizes the error measured by Frobenius norm. Nevertheless, the optimal \( \hat{X} \) cannot be obtained directly by this way as we do not know matrix \( X \) and the proper rank in advance.

Thus we propose to find \( \hat{X} \) as follows:

Objective: \( \min(\text{rank}(\hat{X})) \),

Subject to: \( B \circ \hat{X} = S \).

We make this assumption according to two reasons. On the one hand, since the reconstructed matrix (RM) is generated from the sensory matrix (SM), it is reasonable to be as close as SM. On the other hand, like the environmental matrix (EM), RM should also have a low-rank structure. Given this, it is still difficult to solve this minimization problem because it is non-convex. To bypass this difficulty, we take advantage of the SVD-like factorization, which re-writes Eq. (5) as:

\[
\hat{X} = U \Sigma V^T = LR^T,
\]

where \( L = U \Sigma^{1/2} \) and \( R = V \Sigma^{1/2} \). Substituting Eq. (14) to Eq. (13), we can solve the minimization problem according to the compressive sensing theory in [5], [8]. Specifically, if the restricted isometry property holds [25], minimizing the nuclear norm can result to rank minimization exactly for a low-rank matrix. Hereby, we just need to find matrix \( L \) and \( R \) that minimize the summation of their Frobenius norms:

Objective: \( \min(||L||_F^2 + ||R^T||_F^2) \),

Subject to: \( B \circ (LR^T) = S \).

Looking for \( L \) and \( R \) that strictly satisfy Eq. (15) is likely to fail due to two reasons. First, EMs usually approximate low-rank but not exact low-rank. Second, noises in sensory data may lead to the over-fitting problem if strict satisfaction is required. Thus, instead of solving Eq. (15) directly, we solve the following equation using the Lagrange multiplier method:

\[
\min(||B \circ (LR^T) - S||_F^2 + \lambda(||L||_F^2 + ||R^T||_F^2)),
\]

where the Lagrange multiplier \( \lambda \) allows a tunable tradeoff between rank minimization and accuracy fitness. This solution provides the low-rank approximation but not strict satisfaction.

In Eq. (16), 1) \( B \) and \( S \) are known, 2) any \( || \cdot ||_F^2 \) is non-negative, 3) the optimal values approximate 0 by minimizing all non-negative parts. Hence, \( L \) and \( R \) can be estimated in this optimization problem under the tuning of \( \lambda \).

6.2 Environmental Spatio-Temporal Improvement

ESTI-CS includes two key components: 1) compressive sensing based method for estimating massive missing values and 2) environmental spatio-temporal improvement for increasing the accuracy against diverse loss patterns. On the one hand, the compressive sensing method relies on the low-rank structure. On the other hand, after exploiting the temporal stability and spatial similarity features, we complete ESTI-CS approach by developing Eq. (16) as following:

\[
\min(||B \circ (LR^T) - S||_F^2 + \lambda(||L||_F^2 + ||R^T||_F^2) + ||H(LR^T) - T||_F^2 + ||LR^T \mathbb{T}^T||_F^2),
\]

where \( \mathbb{T} \) and \( \mathbb{T} \) are the spatial and temporal constraint matrices respectively. Three subjects \( ||H(LR^T)||_F^2, ||LR^T \mathbb{T}^T||_F^2 \), and \( ||B \circ (LR^T) - S||_F^2 \) are set to be the same order of magnitude, whose coefficients are 1. Otherwise, they may overshadow the others when solving Eq. (17).

Temporal stability improvement: The temporal constraint matrix \( \mathbb{T} \) captures the temporal stability feature, which outlines that the change between two consecutive time slots is small. Hence, we set \( \mathbb{T} = \)
The Toeplitz matrix defined with central diagonal given by 1, and the first upper diagonal given by -1, and the others given by 0, i.e.,

$$T = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & : \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{t \times t}. \quad (18)$$

This Toeplitz matrix adds the temporal constraint into the estimation. Importing $||LR^T T||_F^2$ into Eq. (17) is equal to induct an additional constraint into the original optimization problem. Since the temporal constraint is an inherent feature of environment, this additional constraint can filter more noises and errors in $LR^T$ estimation.

**Spatial similarity improvement:** The spatial constraint matrix $H$ captures the spatial similarity feature, which reveals that values among one-hop neighbors nodes are usually similar. Hence, we set $H$ to be a row-normalized $H^*$, where $H^* = H + D$. The matrix $H$ is a TM-1H, i.e., the one-hop topology matrix mentioned before. And $D$ is an $n \times n$ diagonal matrix, which is defined with central diagonal given by $\text{diag}(d_1, d_2, \ldots, d_n)$, and the others given by 0. In $D$, $d_i = -\sum H_{(i)}$. For example, if there is a TM-1H:

$$H = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad (19)$$

then,

$$H^* = H + D = \begin{bmatrix} -2 & 1 & 0 & 1 \\ 1 & -3 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & -2 \end{bmatrix}, \quad (20)$$

thus, the corresponding spatial constraint matrix is:

$$H = \begin{bmatrix} 1 & -1/2 & 0 & -1/2 \\ -1/3 & 1 & -1/3 & -1/3 \\ 0 & 1 & -1 & 0 \\ -1/2 & -1/2 & 0 & 1 \end{bmatrix}. \quad (21)$$

The spatial similarity constraint is added by the matrix $H$. Computing the result of $HX$ is to get the differences between the elements and the average value of their one-hop neighbors in $X$. As the same purpose of time improvement part, we introduce the part of minimizing $||H LR^T||_F^2$ into Eq. (17). It takes advantage of the inherent environment feature as an additional constraint in optimization problem, which leads to a more accurate estimation of $LR^T$, i.e., $\hat{X}$.

**Algorithm 1 ESTI-CS algorithm**

**Input:**
- $S_{n \times t}$: sensory matrix
- $B_{n \times t}$: binary index matrix
- $r$: rank bound
- $\lambda$: tradeoff coefficient
- $g$: iteration times

**Output:**
- $\hat{X}_{n \times t}$: estimated environment matrix

**Main procedure:**
1. $L \leftarrow \text{random\_matrix}(n, r)$
2. for $t$ do
3. $R \leftarrow \text{myInverse}(B, L, \lambda, r, S)$
4. $L \leftarrow \text{myInverse}(B^T, R^T, \lambda, r, S^T)$
5. $v \leftarrow \|B \circ (LR^T) - S\|_F^2 + \lambda(\|L\|_F^2 + \|R^T\|_F^2) + \|H LR^T\|_F^2 + \|LR^T T\|_F^2$
6. if $v < \tilde{v}$ then
7. $\hat{L} \leftarrow L; \hat{R} \leftarrow R; \hat{v} \leftarrow v$
8. end if
9. end for
10. $\hat{X} \leftarrow LR^T$
11. return $\hat{X}$

**Procedure** $Y = \text{myInverse}(B, L, \lambda, r, S)$:
1. for $i=1$ to $t$ do
2. $P_i \leftarrow [\text{Diag}(B(i,:))L; \sqrt{\lambda}I_v]$ (19)
3. $Q_i \leftarrow [S(:,i); 0_r]$ (20)
4. $Y(:,i) = (P_i^T P_i) \backslash (P_i^T Q_i)$ (21)
5. end for
6. return $Y$

### 6.3 ESTI-CS Algorithm

We propose an efficient ESTI-CS algorithm to solve the estimation in the optimization problem Eq. (17). The detail pseudo code is shown in Algorithm 1. First, we scale the $T$ and $H$ as all $\|\cdot\|_F$ in Eq. (17) have the same order of magnitude. The scaling method is similar to [31]. Then ESTI-CS algorithm solves the optimization in an iterative manner. $L$ is initialized randomly, so $R$ can be computed by solving the following contradictory equation:

$$B \circ (LR^T) = S \cdot \sqrt{LR^T}. \quad (22)$$

This equation can be rewritten as:

$$\text{Diag}(B(i,:))LR^T_{(i)} = S(i,:), \quad (23)$$

where $i$ ranges from 1 to $t$. This is a combination of multiple standard linear least squares problems. We then have $R^T_{(i)} = (P_i^T P_i) \backslash (P_i^T Q_i)$, where $P_i = [\text{Diag}(B(i,:))L; \sqrt{\lambda}I_v]$ and $Q_i = [S(:,i); 0_r]$. This procedure is reflected by the subfunction $\text{myInverse}$ in the pseudo code. Similarly, once $R^T$ is obtained, $L$ can be re-computed by fixing $R^T$. This mutual re-computing process repeats until the optimal value is reached.

We analyze the complexity of the ESTI-CS algorithm. The key operation is the procedure for computing the inverse matrix, which provides the best approximate solution to the contradictory equation. This pro-
procedure is completed by a matrix multiplication [19]. Thus, its time complexity is $O(rnt)$. Since ESTI-CS repeats the procedure for $q$ times, the total complexity is $O(rntq)$. From our evaluation experience in Section 8, $L$ and $R^T$ usually converge after 5 iterations.

### 6.4 Design Optimization

There are two parameters in ESTI-CS algorithm: rank bound $r$ and tradeoff coefficient $\lambda$, which influence the accuracy of estimated $X$. According to CS theory, the rank of the approximated matrix should be minimized. Therefore, in ESTI-CS, $r$ is the minimal rank between matrix $L$ and $R$, which is also smaller than $n$ or $t$. So we have

$$\text{rank}(\hat{X}) \leq \min(\text{rank}(L), \text{rank}(R)) = r.$$  \hspace{1cm} (24)

The estimation accuracy can be treated as a function of the two parameters, denoted by $f(r, \lambda)$. In order to obtain the optimal parameters, the objective is

$$\min f(r, \lambda).$$  \hspace{1cm} (25)

The error ratio is adopted to indicate the estimation accuracy. Consequently, the function $f()$ builds the bridge between error ratio and the parameters.

We adopt the same genetic algorithm as shown in [19] to derive the optimal rank bound $r$ and tradeoff coefficient $\lambda$. Here, error ratio is served as fitness in the genetic algorithm. The optimal parameters will be obtained when the fitness is stalled after several generations.

The genetic based pseudo code is shown in Algorithm 2. The population is randomly initialized, which will evolve over generations. And the optimal parameters will be selected in the end of this algorithm. In every generation, the population consists of three groups of individuals. The first group includes the kids selected from the last generation based on their fitness. The second group consists of the kids produced by taking the crossover of any two individuals. The third group is produced by the mutation operation. Specifically, we assign a random value to one of parameters within its domain to achieve the mutation. This algorithm iterates until the fitness of the best individual stall.

#### Algorithm 2 Genetic algorithm for optimizing parameters in ESTI-CS

**Input:**
- $l_r, U_r$: lower bound and upper bound of $r$
- $l_\lambda, U_\lambda$: lower bound and upper bound of $\lambda$
- $B_{nst}$: binary index matrix

**Output:**
- Optimal $r$ and $\lambda$

**Main procedure:**
1. $\mathcal{N}(\text{population}) \leftarrow \text{initialize with random numbers uniformly distributed within } [l_r, U_r]$ and $[l_\lambda, U_\lambda]$
2. while (Is stall(fitness)) do
   3. $\mathcal{H} \leftarrow \text{select(}\mathcal{N}\text{)}$
   4. $C \leftarrow \text{crossover(}\mathcal{N}\text{)}$
   5. $M \leftarrow \text{mute(}\mathcal{N}\text{)}$
   6. $\mathcal{N} \leftarrow [\mathcal{H}, C, M]$
3. end while
4. $[r, \lambda] \leftarrow \text{decode (the best kid in } \mathcal{N})$

that approximate the original environmental matrices (EMs) $X_k$.

For simplicity, we study the two-attribute situation as an example. Formally, when $K = 2$, the ERSN problem is formulated as follows: Given $S_1$ and $S_2$, find an optimal solution for $\hat{X}_1$ and $\hat{X}_2$, i.e.,

$$\text{Objective } \min(||\hat{X}_1 - X_1||_F + ||\hat{X}_2 - X_2||_F).$$

Subject to
$$\begin{align*}
B_1 \circ \hat{X}_1 &= S_1, \\
B_2 \circ \hat{X}_2 &= S_2.
\end{align*}$$  \hspace{1cm} (26)

**Normalization:** Since the magnitudes of attributes are different, it may cause one matrix to overshadow another. In order to overcome this issue, $X_1$ and $X_2$ are normalized respectively, i.e., each element is normalized by the maximal element present in the corresponding matrix.

**Low-Rank Matrix Approximation:** Eq.(26) is tied by $X_1$ and $X_2$, so the problem cannot be solved in closure form. However, due to the inherited low-rank feature, this problem can be converted to a rank minimization problem. Thus, the optimal $X_k$ is evaluated by the problem:

$$\min(\text{rank}(\hat{X}_k)), \text{ s.t. } S_k = \hat{X}_k \circ B_k.$$  \hspace{1cm} (27)

Still two problems are up against us: (1) the rank calculating operator $\text{rank}(\cdot)$ is not convex. (2) there is no connection between $X_1$ and $X_2$.

To conquer the difficulty (1), we still utilize SVD-like factorization as $\hat{X} = LR^T$. Thus $\min(\text{rank}(\hat{X}))$ is solvable by looking for $L$ and $R$, which satisfy $\min(\|L\|_F^2 + ||R^T||_F^2)$.

**Compressive Sensing-based Joint Matrix Decomposition:** To overcome the difficulty (2), we need to find the correlation between $X_1$ and $X_2$. Through the JSD analysis in Sec.5.5, we separate the approximation $\hat{X}_1$ and $\hat{X}_2$ by JSD as:

$$\begin{align*}
\hat{X}_1 &= \hat{W} + \Delta_1 \\
\hat{X}_2 &= \hat{W} + \Delta_2
\end{align*}$$  \hspace{1cm} (28)

### 7 Multi-Attribute Component

#### 7.1 MAA Overview

Multi-attribute assistant component can be utilized to improve the accuracy of ESTI-CS when several attributes have correlation. Under such scenario, the proposed ERSN problem is extended to $k$-ERSN problem: Given $K$ sensory matrices (SMs) $S_k$, where $k = 1, 2, \ldots, K$, and these SMs have the same size but different values. The goal is to jointly find the corresponding optimal reconstructed matrices (RMs) $\hat{X}_k$
Since $\hat{W}$, $\hat{\Delta}_1$ and $\hat{\Delta}_2$ inherit the low-rank feature, the $k$-ERSN problem is reformulated as:

$$\text{Objective}\quad \min (||\hat{W}||_* + ||\hat{\Delta}_1||_* + ||\hat{\Delta}_2||_*),$$

$$\text{Subject to}\quad B_1 \circ (\hat{W} + \hat{\Delta}_1) = B_1 \circ X_1, \tag{29}$$

$$B_2 \circ (\hat{W} + \hat{\Delta}_2) = B_2 \circ X_2.$$  

where $|| \cdot ||_*$ is the nuclear norm which is defined as the sum of singular values, e.g., $||X||_* = \sum_{i=1}^r \sigma_i(X)$.

Furthermore, using SVD-like factorization, $||\hat{W}||_* + ||\hat{\Delta}_1||_* + ||\hat{\Delta}_2||_*$ in Eq.(29) is rewritten as:

$$||L_W||_F^2 + ||R_W||_F^2 + ||L_1||_F^2 + ||R_1||_F^2 + ||L_2||_F^2 + ||R_2||_F^2.$$  

where $L_W, L_1, L_2$ are $n \times r$ matrices and $R_W, R_1, R_2$ are $r \times m$ matrices. Moreover, $W = L_2 R_2^T, \Delta_1 = L_1 R_1^T$ and $\Delta_2 = L_2 R_2^T$. For short, Eq.(30) is denoted by $\sum ||L_j||_F^2 + \sum ||R_j||_F^2$, where $j = 1, 2, W$.

To avoid overfitting, the $k$-ERSN problem is rewritten to be a non-stationary optimization problem using the Lagrange multiplier method, i.e.,

$$\min (\lambda (\sum ||L_j||_F^2 + \sum ||R_j||_F^2)$$

$$+ ||B_1 \circ (L_W R_W^T + L_1 R_1^T) - S_1||_F^2)$$

$$+ ||B_2 \circ (L_W R_W^T + L_2 R_2^T) - S_2||_F^2), \tag{31}$$

Eq.(31) is the core of MAA component, which is solvable because (1) $B_1, B_2, S_1$ and $S_2$ are known, (2) each $|| \cdot ||_F^2$ is non-negative, (3) the optimal value can be reached by minimizing all non-negative parts to zero. Hence, $X_1$ and $X_2$ can be estimated by combining Eq.(31) and Eq.(28).

**ESTI-CS with MAA** is to reconstruct several (two as example) environments according to

$$\min (\lambda (\sum ||L_j||_F^2 + \sum ||R_j||_F^2)$$

$$+ ||B_1 \circ (L_W R_W^T + L_1 R_1^T) - S_1||_F^2)$$

$$+ ||B_2 \circ (L_W R_W^T + L_2 R_2^T) - S_2||_F^2) \tag{32}$$

$$(\sum ||L_j||_F^2 + \sum ||R_j||_F^2)$$

It can be seen that Eq.(32) is the combination of Eq.(31) and Eq.(17). The first three items of Eq.(32) utilize the low-rank feature, which is the fundamental compressive sensing, the fourth and fifth items incorporate the spatial similarity improvement, the last two items merge the temporal stability improvement, and the multi-attribute assistant component is added a lastly into Eq.(32) by $L_j$ and $R_j$, where $j = 1, 2, W$.

**Extension:** The MAA component is also suitable for the case of more attributes. For instance, if we obtain $k$ attributes in one WSN, represented by $X_1, X_2, \cdots, X_k$. The utilization of MAA is to rewrite Eq.(29) into

$$||\hat{W}||_* + ||\hat{\Delta}_1||_* + ||\hat{\Delta}_2||_* + \cdots + ||\hat{\Delta}_k||_*.$$  

Then the $k$-ERSN problem can be solved by the similar method of the above two-attribute case.

### Algorithm 3 MAA Algorithm

**Input:**

- $S_1$ and $S_2$: sensory matrices sets
- $r$: rank approximation
- $\lambda$: tradeoff coefficient
- $g$: iteration times

**Output:**

- $\hat{X}_1$ and $\hat{X}_2$: estimated environment matrices

**Main Procedure:**

1. **Normalization**
   - $\alpha_1 \leftarrow \max(S_1); \quad \alpha_2 \leftarrow \max(S_2)$
   - $S_1 \leftarrow S_1 / \alpha_1; \quad S_2 \leftarrow S_2 / \alpha_2$

2. **Approximation**
   - $L_W \leftarrow \text{rand}(n, r)$
   - $L_1 \leftarrow \text{rand}(n, r)$
   - $L_2 \leftarrow \text{rand}(n, r)$
   - $R_1 \leftarrow \text{rand}(r, t)$
   - $R_2 \leftarrow \text{rand}(r, t)$

3. for $1 \leq g$ do
   - $F_1 = S_1 - B_1 \circ (L_1 R_1^T)$
   - $F_2 = S_2 - B_2 \circ (L_2 R_2^T)$
   - $W = \text{crossInverse}(B_1, B_2, L_W, L_1, R_1, F_1, F_2)$
   - $L_1 = \text{singleInverse}(B_1 R_1, L_1, \lambda, r, F_1)$
   - $L_2 = \text{singleInverse}(B_2 R_2, L_2, \lambda, r, F_2)$
   - $L_W = \text{crossInverse}(B_1, B_2, L_W, L_1, R_1, F_1, F_2)$
   - $L_1 = \text{singleInverse}(B_1 R_1, L_1, \lambda, r, F_1)$
   - $L_2 = \text{singleInverse}(B_2 R_2, L_2, \lambda, r, F_2)$

4. $v \leftarrow \text{Eq.(31)}$

5. if $v < \hat{v}$ then
   - $\hat{L}_W \leftarrow L_W$; $\hat{L}_1 \leftarrow L_1$; $\hat{R}_1 \leftarrow R_1$; $\hat{L}_2 \leftarrow L_2$; $\hat{R}_2 \leftarrow R_2$; $\hat{v} \leftarrow v$

6. end if

7. end for

8. $X_1 \leftarrow \alpha_1 (L_W R_W^T + \hat{L}_1 R_1^T)$
9. $X_2 \leftarrow \alpha_2 (L_W R_W^T + \hat{L}_2 R_2^T)$
10. return $(\hat{X}_1, \hat{X}_2)$

**Procedure $Y = \text{singleInverse}(B, L, \lambda, r, S)$:**

1. for $i=1$ to $t$ do
   - $P_i \leftarrow [B L (:, i); \sqrt{\lambda} I_r]$ 
   - $Q_i \leftarrow [S (:, i); 0_r]$ 
   - $Y (:, i) = (P_i P_i^T) \backslash (P_i^T Q_i)$
2. end for
3. return $Y$

**Procedure $Y = \text{crossInverse}(B_1, B_2, L, \lambda, r, S_1, S_2)$:**

1. for $i=1$ to $t$ do
   - $P_i \leftarrow [B_1 L (:, i); B_2 L (:, i); \sqrt{\lambda} \ast I_r]$ 
   - $Q_i \leftarrow [S_1 (:, i); S_2 (:, i); 0_r]$ 
   - $Y (:, i) = (P_i P_i^T) \backslash (P_i^T Q_i)$
2. end for
3. return $Y$

### 7.2 MAA Algorithm

Algorithm 3 shows the pseudo-code of optimization algorithm with the MAA component. Note that we only present the core of MAA component in this section, which is to solve Eq.(31). And the realization of spatial-temporal improvement in Eq.(32) is the same method in ESTI-CS, we do not repeat it here.

The algorithm solves the problem in an iterative manner. First, $L_1, L_2, L_W, R_1$, and $R_2$ matrices are initialized randomly. Then, $R_W$ can be calculated from
the initialized matrices by solving the equation:
\[
\begin{bmatrix}
    B_1 \circ (L_W R_W^I) \\
    B_2 \circ (L_W R_W^I) \\
    \sqrt{\lambda} R_W^I
\end{bmatrix} = \begin{bmatrix}
    S_1 - L_1 R_1^I \\
    S_2 - L_2 R_2^I \\
    0
\end{bmatrix}.
\]

Eq.(34) is solvable using the linear least square method. The subfunction \texttt{crossInverse} in Algorithm 3 shows how to compute \(R_W\). After \(R_W\) is obtained, \(L_W\) can be computed using the same procedure by fixing \(R_W\).

Similarity, any of \(L_1, L_2, R_1\) and \(R_2\) is computed by fixing the other three. Using the iterative manner, all these four matrices can be obtained one-by-one. The pseudo code is shown in the subfunction \texttt{singleInverse}.

The analyses of rank approximation \(r\), the tradeoff coefficient \(\lambda\), and the computational complexity are the same as Algorithm 1.

8 PERFORMANCE EVALUATION

8.1 Methodology

The proposed ESTI-CS approach is compared with existing algorithms for missing data interpolation for environmental reconstruction in WSNs.

**Ground truth.** Since the performance evaluation needs complete EMs \(X\) to compute the metric of error ratio (ER), we utilize the datasets as shown in Tab. 2. Six EMs are adopted: indoor-temp, indoor-light, forest-temp, forest-light, ocean-temp and ocean-light.

**Methods.** To verify the effectiveness of ESTI-CS, four classic interpolation methods are selected for comparison. They are compressive sensing (CS) [19] with time complexity \(O(rnt_0)\), Delaunay Triangulation (DT) [13] with time complexity \(O(nt \log nt)\), Multi-channel Singular Spectrum Analysis (MSSA) [32] with time complexity \(O(rnt \log nt + r^2nt)\), and K-Nearest Neighbor (KNN) [7] with time complexity \(O(nt)\). The parameter \(K\) in KNN is set to be \(\sum_{i=1}^n H_i/\bar{H}\). The parameter \(M\) in MSSA is set to 32 as suggested by [32].

**Procedure.** The procedure of simulation is: 1) Generate BIM \(B\) according to four loss patterns. 2) Compute SM \(S\) according to Definition 3: \(S = B \circ X\). 3) All interpolation algorithms being tested take SMs as input and generate RMs. 4) The accuracy metric ER is computed for the difference between EMs and RMs. And finally, these errors are compared for performance evaluation.

**Series.** Three series of experiments are evaluated. The basic experiment measures the performance of different algorithms against typical random loss probability. The second experiment evaluates the performance in diverse data loss patterns. And the third experiment compares the performance of ESTI-CS with MAA and ESTI-CS without MAA.

8.2 Performance Analysis: Basic Comparison of Algorithms in Random Loss Pattern

In the basic comparison, we test the error ratios under diverse algorithms on the element random loss (ERL) pattern only. The data loss rate \(p_{\text{ERL}}\) ranges from 10% to 90%. If the loss rate is 0, i.e., the dataset is complete, it is unnecessary to be interpolated. If the loss rate is raised to 100%, i.e., all data are lost, no methods can recover the EMs.

Fig. 5 shows the comparison results. The X-axis presents the data loss probability, and the Y-axis is the value of ER, which represents the reconstruction accuracy. In general, ER increases with the data loss rate.

In the indoor-temp Fig. 5(a), ESTI-CS shows the best performance. Even 90% data have been lost, ESTI-CS still can reconstruct the environment with \(ER \leq 10\%\). While ER of CS is about 19%, DT is close to 38%, and ER of KNN and MSSA are more than 60%. ESTI-CS is much better than other algorithms in this scenario. In the indoor-light Fig. 5(b), ESTI-CS still outperforms the others, but the advantage is less significant than that in indoor-temp. The reason is that the indoor temperature change has strong spatio-temporal feature. However, the change of indoor light is largely influenced by the light switch. So the indoor light dataset observes more artificial changes than spatio-temporal stability.

The performance of Forest-Temp and the Forst-Light are similar. The reason is that both the temperature and the light are mainly effected by the sun due to an outdoor application. These two environment attributes have strong correlation. As shown in Fig. 5(c) and Fig. 5(d), ESTI-CS achieves the best environment reconstruction among the five algorithms. CS, MSSA and DT fall behind ESTI-CS a little. KNN is not bad when \(p_{\text{ERL}} < 50\%\), but when \(p_{\text{ERL}} > 50\%\), ER of KNN increases quickly.

In the ocean-temp Fig. 5(e), ESTI-CS and DT produce the similar performance. When the data loss is 90%, they achieve ER< 30%. Meanwhile, the ERs of CS, KNN and MSSA are bigger. In the ocean-light Fig. 5(f), the performance of ESTI-CS and DT are similar with the range of loss rate from 10% to 80%. When the loss rate increases to 90%, ER of DT also increases rapidly, and ER of ESTI-CS still keeps within 22%. These two figures indicate that ESTI-CS perform better than DT, CS, KNN and MSSA in this outdoor and small-scale WSN scenario.

Overall, ESTI-CS obtain lower interpolation error, which can be used in almost all tested datasets with different loss ratios. KNN and DT produce similar but the poor ER performance, because both of them interpolate with only the space relation among nodes but no time relation consideration. CS and MSSA are better than KNN and DT, but still worse than ESTI-CS. Especially, at the high data loss cases (data
loss $\geq 80\%$), ESTI-CS exhibits an evident advantage over other algorithms. In all dataset, ESTI-CS can successfully achieve an environment reconstruction with 20% error when there are 90% data are missing.

### 8.3 Performance Analysis: Data Loss Patterns Comparison

In Fig. 6, we plot the comparison histograms of five algorithms for reconstructing the environment with four different data loss patterns.

In the simulation for BRL pattern, each of the six EMs is set to lose data with the block pattern as shown in Fig. 1(b). The scale and the number of the blocks are random, but the amount of total data loss is 50% in this simulation. In Fig. 6(a), most algorithms in most EMs perform not well. For example, in forest-light, ER of all algorithms are bigger than 60%. The reasons are 1) in the forest, many shadows disturb the spatio-temporal stability. 2) if large blocks of data lose, spatio-temporal optimized estimation is helpless either. These two reasons lead to the result. However, in indoor-temp, ocean-temp, and ocean-light, the environment changes are smoothly, ER of ESTI-CS are less than 5% despite 50% BRL data loss. Even indoor-light, forest environments, ESTI-CS is still a bit better than the others. In addition, we find that KNN is in big trouble for estimating the missing data in BRL.

In the simulation of EFLR pattern, the rows are randomly selected, and then all elements after the starting points in those rows are lost as shown in Fig. 1(d). The total amount of data loss is set to 50%. The results of all algorithms are between those in EFLR and BRL. For ESTI-CS, these results are reasonable. The reason is that ESTI-CS can only use space optimization, but the time optimization has no effect due to elements lost in all time of a node. Note that all algorithm plays not well for ocean-light in this simulation. Since the scale of ocean-light is small, after some additional rows are lost, it becomes smaller, which is hard to be estimated.

In summary, ESTI-CS outperforms CS, KNN, DT and MSSA in any dataset in the combinational loss pattern.

In the simulation of SELR pattern, the starting points are randomly selected, and then all elements after the starting points in those rows are lost as shown in Fig. 1(d). The total amount of data loss is set to be 50%. The results of all algorithms are between those in EFLR and BRL. For ESTI-CS, these results are reasonable. The reason is that ESTI-CS can only use space optimization, but the time optimization has no effect due to elements lost in all time of a node. Note that all algorithm plays not well for ocean-light in this simulation. Since the scale of ocean-light is small, after some additional rows are lost, it becomes smaller, which is hard to be estimated.

### 8.4 Performance Analysis: ESTI-CS with MAA

In this experiment, we evaluate the benefit from MAA component for ESTI-CS. The data loss pattern and loss rate setting are the same as the basic comparison in Sec.8.2.

In Fig. 7, we illustrate the ER results of ESTI-CS-MAA algorithm and compare them with ESTI-CS in the case of two attributes under random loss pattern. In every dataset, two attributes temperature and light
are reconstructed together by ESTI-CS-MAA. It can be seen that reconstruction accuracy of ESTI-CS-MAA is universally better than that of ESTI-CS in three real WSN datasets.

Fig. 7(a) and Fig. 7(b) show ESTI-CS-MAA is slightly better than ESTI-CS in Intel Indoor dataset. In face of 90% data loss, ESTI-CS-MAA improves the ER by 2% in Indoor-Temp shown in Fig. 7(a) and 3% shown in Indoor-Light in Fig. 7(b) compared with ESTI-CS.

In GreenOrbs dataset, ESTI-CS-MAA performs better than ESTI-CS. Also in the case of 90% data loss, The MAA-enabled algorithm outperforms baseline ESTI-CS by 7% in Forest-Temp shown in Fig. 7(c) and 6% in Forest-Light shown in Fig. 7(d).

The results of ESTI-CS-MAA are significantly better than those of ESTI-CS in OceanSense dataset. As shown in Fig. 7(e) and Fig. 7(f), 14% and 12% ER are improved by ESTI-CS-MAA respectively in Ocean-Temp and in Ocean-Light.

Recall the correlation analysis of Fig. 4, the correlation between temperature and light in OceanSense is higher than that in GreenOrbs and even higher than that in Intel Indoor. We summarize that MAA can further improve the ESTI-CS on reconstruction accuracy when multi-attribute correlation exists. Moreover, the higher the correlation is, the more improvement the MAA provides.

9 CONCLUSION

In this paper, we studied the data loss and reconstruction problem in WSNs. We verified the massive data loss in real datasets and modeled the special data loss patterns of WSNs. Then, we mined the low-rank, spatial, temporal, and correlation features from WSN datasets. By drawing on these observations, we designed ESTI-CS with MAA algorithm to estimate the missing data. The proposed algorithm combines the benefits of compressive sensing, environmental space-time, and multi-attribute correlation features. Trace-driven experiments illustrated that ESTI-CS outperforms existing interpolation methods.

There are two avenues for our future work. First, study the tradeoff between the computation time and accuracy in environment reconstruction. Second, data interpolation in mobile sensor networks is also interesting and challenging.

ACKNOWLEDGMENT

This research was supported by NSFC under grant No. 61303202, No. 61073158, No. 61100210, STCSM Project No. 12dz1507400, Doctoral Program Foundation under grant No. 2011073120021, FQRNT grant
Fig. 7. Error ratio performance of ESTI-CS with MAA in the basic data loss pattern: element random loss.
Linghe Kong (S’09-M’13) is currently a research assistant professor in Shanghai Jiao Tong University and a postdoctoral researcher in Singapore University of Technology and Design. He received his PhD degree in Computer Science with Shanghai Jiao Tong University 2012, Dipl. Ing. degree in Telecommunication with TELECOM Sud-Paris 2007, B. E. degree in Automation with Xidian University 2005. His research interests include wireless sensor networks and mobile computing.

Mingyuan Xia received the B.E. degree in Computer Science and Engineer from Shanghai Jiao Tong University, China, in 2011, and is now a PhD candidate in the School of Computer Science at McGill University, Canada. His research interests include wireless sensor networks and operation systems.

Xiao-Yang Liu received his B. E. degree in computer science in Huazhong University of Science and Technology, China, in 2010. He is currently working toward the PhD degree in Department of Computer Science and Engineer in Shanghai Jiao Tong University. His research interests are in the area of wireless sensor networks.

Guangshuo Chen received the B.E. degree in Electronic and Information Engineering from Shanghai Jiao Tong University, China, in 2011, and is now an academic Master degree candidate in Computer Science from Shanghai Jiao Tong University. His research interests include Wireless Sensor Networks (WSNs), Body Area Networks and Vehicular Networks. At present, he focus on the data recovery problem in WSNs.

Min-You Wu (S’84-M’85-SM’96) received the M.S. degree from the Graduate School of Academia Sinica, Beijing, China, in 1981 and the Ph.D. degree from Santa Clara University, Santa Clara, CA in 1984. He is a professor in the Department of Computer Science and Engineering at Shanghai Jiao Tong University and a research professor with the University of New Mexico. He serves as the Chief Scientist at Grid Center of Shanghai Jiao Tong University. His research interests include grid computing, wireless networks, sensor networks, multimedia networking, parallel and distributed systems, and compilers for parallel computers.

Xue Liu received the B.S. degree in mathematics and the MS degree in automatic control both from Tsinghua University, China, and the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2006. He is an associate professor in the School of Computer Science at McGill University. He has also worked as the Samuel R. Thompson associate professor in the University of Nebraska-Lincoln and HP Labs in Palo Alto, California. His research interests include computer networks and communications, smart grid, real-time and embedded systems, cyber-physical systems, data centers, and software reliability.