Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 13(2): 105-112, Jun. 2015 **Regular paper**

A New Estimation Model for Wireless Sensor Networks Based on the Spatial-Temporal Correlation Analysis

Xiaojun Ren, HyonTai Sug*, and HoonJae Lee, *Member, KIICE*

Department of Ubiquitous IT, Dongseo University, Busan 617-716, Korea

Abstract

The estimation of missing sensor values is an important problem in sensor network applications, but the existing approaches have some limitations, such as the limitations of application scope and estimation accuracy. Therefore, in this paper, we propose a new estimation model based on a spatial-temporal correlation analysis (STCAM). STCAM can make full use of spatial and temporal correlations and can recognize whether the sensor parameters have a spatial correlation or a temporal correlation, and whether the missing sensor data are continuous. According to the recognition results, STCAM can choose one of the most suitable algorithms from among linear interpolation algorithm of temporal correlation analysis (TCA-LI), multiple regression algorithm of temporal correlation analysis (TCA-MR), spatial correlation analysis (SCA), spatial-temporal correlation analysis (STCA) to estimate the missing sensor data. STCAM was evaluated over Intel lab dataset and a traffic dataset, and the simulation experiment results show that STCAM has good estimation accuracy.

Index Terms: Data mining, DESM, Missing sensor data, STCAM

I. INTRODUCTION

With the rapid development of wireless communication, microelectronics, and embedded computing technologies, sensor networks are widely used in certain fields, such as the military, environment, and medicine. Therefore, nowadays, these networks have become a popular topic of research. In a wireless sensor network, sensors always communicate with the server and other sensors (e.g., for sending data or accepting data). However, in the process of communication, we can expect the transmitted sensor data to get lost or corrupted for many reasons, such as bad weather conditions, the sensor node's communication ability, wireless signal strength, power outage at the sensor node, or a relatively high bit error rate of the wireless radio transmissions as compared with wired communications. In general, we can re-query data or discard data, but re-querying data is a naive alternative as it may induce a long wait or quicken the power exhaustion of the node, and most importantly, it does not guarantee having the original reading available. Discarding data is also a bad choice as it may lead to the loss of some interesting data. Therefore, it is essential to develop a technique for estimating missing data.

Data mining can produce knowledge from the existing data, and this knowledge can be used for estimating the missing sensor data. However, the existing missing sensor data estimation approaches do not achieve good results (as discussed in the following section). Therefore, in this paper, we propose a new estimation model based on a spatial-temporal correlation analysis (STCA). This model can discover intrinsic relationships among sensors and then incorporate the intrinsic relationships and the spatial-temporal

Received 04 February 2015, **Revised** 01 April 2015, **Accepted** 01 May 2015
***Corresponding Author** HyonTai Sug (E-mail: sht@dongseo.ac.kr, Tel: +82-51-320-1733)
Department of Ubiquitous IT, Dongseo University, 47 Jurye-ro, Sasang-gu, Busan 617-716, Korea.

Open Access **http://dx.doi.org/10.6109/jicce.2015.13.2.105** print ISSN: 2234-8255 online ISSN: 2234-8883

relationship into data estimation. Finally, STCAM is tested with data from a traffic monitoring sensor application.

II. RELATED WORKS

In fact, the topic of missing data estimation belongs to the field of statistics, and many researchers have conducted a considerable amount of research on this topic by using methods such as mean substitution, linear regression, Bayesian estimation, expectation maximization, k-nearest neighbor, and neural networks [1, 2]. However, because of the characteristics of a wireless sensor network, these techniques cannot provide a good estimation of the missing sensor data. To solve the problem of missing sensor data, many techniques have been proposed.

To avoid the problem of missing sensor data, many researchers have redesigned the sensor network architecture. NASA/JPL [3] is one of the most famous architectures. In NASA/JPL, if one sensor fails, its neighboring sensors compensate for the lost data by increasing their sampling rates. This implies that there must be a tight collaboration among sensors for a sensor to know that its neighboring sensor has failed. This increases the power consumption of every sensor even during its normal operation. Further, this approach does not address how sampling rates should be adjusted in order to guarantee good QoS. It is also possible that when some neighboring sensors fail, no sampling adjustment can potentially compensate for the missing values.

Some of the researchers used association rule mining to estimate the missing sensor data. Halatchev and Gruenwald [4] proposed the WARM algorithm. In this algorithm, if sensor node *a* fails, WARM will find its neighbor sensor node *b* and use *b*'s data to estimate *a*'s missing data. WARM makes use of the sliding window concept, where only the latest *w* rounds of data reports are stored and used for estimation. However, the algorithm has one limitation, which is its disregard of the temporal aspect since it views all data as equally important. Gruenwald et al. [5] proposed an improved algorithm called FARM, which uses association rule mining to discover intrinsic relationships among sensors and incorporates them into the data estimation while taking data freshness into consideration. However, WARM and FARM can only be used in the case of discrete data; because most of the sensor data are numeric, WARM and FARM cannot be used widely.

III. ESTIMATION MODEL BASED ON SPATIAL-TEMPORAL CORRELATION ANALYSIS

In fact, there are two types of missing sensor data, namely single missing data elements and continuous missing data; therefore, STCAM must have the ability to provide different solutions for different types of missing data according to the sensor's spatial-temporal correlation. Before the given STCA, we will discuss the problem description, temporal correlation algorithm (TCA), spatial correlation algorithm (SA), and spatial-temporal correlation algorithm (STCA).

A. Problem Description

STCAM uses a temporal series form to represent the collected data of a sensor node a_k. The temporal series form is as follows:

$$S_k = (<V_{k1}, T_1>, <V_{k2}, T_2>, <V_{k3}, T_3> ..... <V_{kn}, T_n>),$$

where $T_1.....T_n \in R$ denote the sampling time and $V_{k1}.....V_{kn} \in R$ represent the sampling values of sensor node a_k at time $T_1.....T_n$. Assuming that V_{ki} denotes the missing sensor data and V_{ki}' represents the estimated sensor data at time T_i, we can reduce the problem of the estimation of the missing sensor data to the calculation of the smallest value of $|V_{ki}' - V_{ki}|$.

B. Temporal Correlation Algorithm

In some applications, the data of the monitoring parameter have a tight temporal correlation, such as temperature, humidity, and light intensity. Therefore, we can use temporal correlations to build the TCA model. In the next section, we will introduce two algorithms, namely the linear interpolation algorithm (TCA-LI) and multiple regression algorithm (TCA-MR).

1) TCA-LI Algorithm

The linear interpolation algorithm is a method of curve fitting using linear polynomials, which have a high efficiency. In this section, TCA-LI can be expressed by the following formula [6]:

$$V_{ki}' = V_{iu} + \frac{T_i - T_u}{T_v - T_u}(V_{iv} - V_{iu}), \quad (1)$$

where T_u and T_v denote the two nearest time points from T_i, and $T_u < T_i < T_v$; V_{ki}' denotes the estimated sensor data at time T_i, and V_{iu} and V_{iv} represent the sampling data at time T_u and T_v, respectively.

For a single missing data element, the TCA-LI algorithm can give a better attestation value, but if the missed sensor data are continuous, the accuracy of the TCA-LI algorithm decreases, as shown in Fig. 1. Sensor V measures the temperature every 24 minutes. Assuming that T_{1176} (V_{1176} = 32.90) is missed and that T_{1152} (V_{1150} = 33.10) and T_{1200} (V_{1200} = 32.50) are the two nearest time points from T_{1176}, we find that V_{1176}' = 32.80 is close to V_{1176}. However, assuming the data between T_{1008} and T_{1272}, we find that

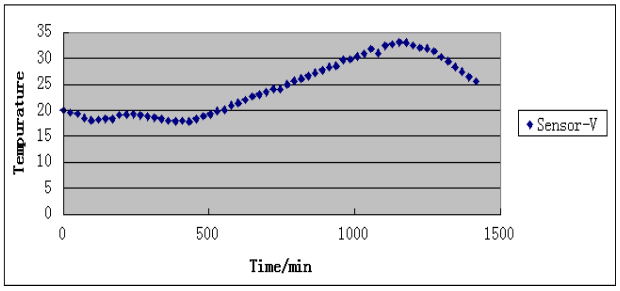

Fig. 1. Temperature data collected by sensor V for one day.

T_{984} (V_{984} = 29.80) and T_{1296} (V_{1296} = 30.30) are the two nearest time points from T_{1176}; therefore, V'_{1176} = 29.05 and the value of $|V'_{1176} - V_{1176}|$ is very large. Hence, the TCA-LI algorithm is only used for estimating single missing data elements.

2) TCA-MR Algorithm

From the above section, we see that TCA-LI has good accuracy for single missing sensor data elements in TCA, but for continuous missing sensor data, TCA-LI cannot provide good estimation data. Therefore, in this section, we will introduce the multiple regression algorithm (TCA-MR) to estimate the continuous missing sensor data of the TCA model. Assuming that V_{ki} denotes the missing data of sensor node a_k at time T_i, the problem of estimating V_{ki}' can be solved by using the following multiple regression formula:

$$V_{ki}' = \beta_0 + \beta_1 V_{k(i-1)} + \beta_2 V_{k(i-2)} + ... + \beta_m V_{(i-m)}, \quad (2)$$

where $\{\beta_0, \beta_1, \beta_2 ..... \beta_m\}$ denote regression coefficients, which represent the contribution level for V_{ki}'.

To estimate V_{ki}', we should use the training dataset to estimate the value of $\{\beta_0, \beta_1, \beta_2 ..... \beta_m\}$. Assuming that the training dataset is $\{V_{ki}, V_{k(i+1)}, V_{k(i+2)} ..... V_{kj}\}$, $j > i + 2m + 1$. To estimate $\{\beta_0, \beta_1, \beta_2 ..... \beta_m\}$, we should build h linear equations ($h > m + 1$) that can be expressed as follows:

$$\begin{cases} V_{kj} = \beta_0 + \beta_1 V_{k(j-1)} + \beta_2 V_{k(j-2)} + \ldots + \beta_m V_{k(k-m)} \\ V_{k(j-1)} = \beta_0 + \beta_1 V_{k(j-2)} + \beta_2 V_{k(j-3)} + \ldots + \beta_m V_{k(j-m-1)} \\ \cdots\cdots\cdots \\ V_{k(j-h)} = \beta_0 + \beta_1 V_{k(j-h-1)} + \beta_2 V_{k(j-h-2)} + \ldots + \beta_m V_{k(j-h-m)} \end{cases} . \quad (3)$$

Let

$$V = (V_{kj}, V_{k(j-1)} ..... V_{k(j-h)})^T ,$$

$$\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_m)^T ,$$

$$X = \begin{bmatrix} 1 & V_{k(j-1)} & V_{k(j-2)} & \cdots & V_{k(j-m)} \\ 1 & V_{k(j-2)} & V_{k(j-3)} & \cdots & V_{k(j-m-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & V_{k(j-h-1)} & V_{k(j-h-2)} & \cdots & V_{k(j-h-m)} \end{bmatrix} .$$

Therefore, Eq. (3) can be rewritten as follows:

$$V = X\beta . \quad (4)$$

The coefficient β can be estimated by using the least-squares approach [7], which can be expressed as follows:

$$\beta = (X^T X)^{-1} (X^T V) . \quad (5)$$

After we calculate the value of coefficient β, we can use Equation (2) to estimate the continuous missing sensor data of TCA.

C. SCA Algorithm

For continuous missing data and loose temporal correlation parameters, the TCA algorithm cannot provide a good estimation value for the missing data. However, the SCA algorithm can discover the spatial relationship between the sensor nodes and use the discovered spatial knowledge to estimate the missing data.

Assuming that V_{ki} denotes the missing data of sensor node a_k at time T_i and $\{a_1, a_2 ..... a_m\}$ represent the neighbors of a_k, we find that $\{V_{1i}, V_{2i} ..... V_{mi}\}$ represent the data values of $\{a_1, a_2 ..... a_m\}$ at time T_i. The problem of estimating V_{ki}' can be solved by $\{V_{1i}, V_{2i} ..... V_{mi}\}$ using the multiple regression as follows:

$$V_{ki}' = \beta_0 + \beta_1 V_{1i} + \beta_2 V_{2i} + ... + \beta_m V_{mi} , \quad (6)$$

where $\{\beta_0, \beta_1, \beta_2 ..... \beta_m\}$ denote the regression coefficients, which represent the contribution level of V_{ki}'.

To calculate V_{ki}', SCA needs a dataset to estimate the value of $\{\beta_0, \beta_1, \beta_2 ..... \beta_m\}$. According to the solution rules of linear equations, the dataset contains at least (m + 1) groups of $\{V_1, V_2 ..... V_m\}$. Note that when $h > m + 1$, the linear equations can be expressed as follows:

$$\begin{cases} V_{k1} = \beta_0 + \beta_1 V_{11} + \beta_2 V_{21} + ... + \beta_m V_{m1} \\ V_{k2} = \beta_0 + \beta_1 V_{12} + \beta_2 V_{22} + ... + \beta_m V_{m2} \\ ..... \\ V_{kh} = \beta_0 + \beta_1 V_{1h} + \beta_2 V_{2h} + ... + \beta_m V_{mh} \end{cases} . \quad (7)$$

Let

$$V = (V_{k1}, V_{k2} ..... V_{kk})^T ,$$

$$\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_m)^T ,$$

$$X = \begin{bmatrix} 1 & V_{11} & V_{21} & \cdots & V_{m1} \\ 1 & V_{12} & V_{22} & \cdots & V_{m2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & V_{1h} & V_{2h} & \cdots & V_{mh} \end{bmatrix} .$$

Therefore, Eq. (3) can be represented by using matrix algebra as follows:

$$V = X\beta \ . \tag{8}$$

Hence, we can calculate the value of coefficient β as follows:

$$\beta = (X^T X)^{-1}(X^T V) \ . \tag{9}$$

D. STCA Algorithm

The TCA algorithm is always used for estimating tight temporal correlations and single missing data elements. The SRA algorithm is used for estimating tight spatial correlations. However, when the temporal or spatial correlation is unknown, TCA or SCA may not give a good estimation value. To solve this problem, the STCA algorithm is proposed. This algorithm takes into account the weight of the temporal and spatial correlations; therefore, the STCA algorithm can be represented as follows:

$$V_{ki}^{'} = W_s V_{ki}^{'S} + W_T V_{ki}^{'T} \ ,$$

where W_s and W_T denote the weight of $V_{ki}^{'s}$ and $V_{ki}^{'T}$, respectively; $W_s + W_T = 1$; $V_{ki}^{'s}$ represents the result of the SCA algorithm, and $V_{ki}^{'T}$ denotes the result of the TCA algorithm.

To obtain the optimum value of W_s and W_T, STCA calculates the residual sum of squares (RSS) as follows:

$$\begin{aligned}
RSS &= \sum_{i=1}^{h}(V_{ki} - V_{ki}^{'})^2 = \\
&\sum_{i=1}^{h}(W_s V_{ki} - W_s V_{ki}^{'s} + W_T V_{ki} - W_T V_{ki}^{'T})^2 = \\
&\sum_{i=1}^{h}(W_s (V_{ki} - V_{ki}^{'s}) + W_T (V_{ki} - V_{ki}^{'T}))^2 = \\
&\sum_{i=1}^{h}(W_s e_{ki}^{s} + W_T e_{ki}^{T})^2 = \\
&\sum_{i=1}^{h}([W_s \quad W_T][e_{ki}^{s} \quad e_{ki}^{T}]^T)^2 = \\
&\sum_{i=1}^{h}\left([W_s \quad W_T]\begin{bmatrix}(e_{ki}^{s})^2 & e_{ki}^{s}e_{ki}^{T} \\ e_{ki}^{s}e_{ki}^{T} & (e_{ki}^{T})^2\end{bmatrix}[W_s \quad W_T]^T\right)^2 = \\
&[W_s \quad W_T]\begin{bmatrix}\sum_{i=1}^{h}(e_{ki}^{s})^2 & \sum_{i=1}^{h}e_{ki}^{s}e_{ki}^{T} \\ \sum_{i=1}^{h}e_{ki}^{s}e_{ki}^{T} & \sum_{i=1}^{h}(e_{ki}^{T})^2\end{bmatrix}[W_s \quad W_T]^T
\end{aligned} \ , \tag{10}$$

where e_{ki}^{s} and e_{ki}^{T} denote the estimation error of $V_{ki}^{'s}$ and $V_{ki}^{'T}$, respectively, and h represents the number of selected datasets.

Let

$$E = \begin{bmatrix}\sum_{i=1}^{h}(e_{ki}^{s})^2 & \sum_{i=1}^{h}e_{ki}^{s}e_{ki}^{T} \\ \sum_{i=1}^{h}e_{ki}^{s}e_{ki}^{T} & \sum_{i=1}^{h}(e_{ki}^{T})^2\end{bmatrix},$$

$$W = [W_s \quad W_T].$$

Therefore, this question of getting optimum value of W_s and W_T becomes a quadratic programming problem:

$$Min(RSS) = WEW^T \tag{11}$$

Therefore, we can use the least-squares approach [7] to obtain an optimal solution as follows:

$$[W_s \quad W_T] = \frac{E^{-1}X^T}{X^T E^{-1} X} \ ,$$

where

$$X = [1 \quad 1] \ .$$

E. Correlation Analysis Algorithm

We use Pearson's product-moment correlation coefficient to measure the correlation of the output variable and the input variable. The value of ρ is between +1 and −1 (inclusive), where 1 denotes a total positive correlation, 0 represents no correlation, and −1 indicates a total negative correlation. The formula for ρ is as follows:

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \ , \tag{12}$$

where y denotes the output variable; x represents the input variable; μ_x and μ_y denote the mean of x and y, respectively; and σ_x and σ_y indicate the standard deviation of x and y, respectively.

Further, $0.5 < |\rho| \le 1$ is regarded as a high correlation, $0.3 < |\rho| \le 0.5$ as a medium correlation, and $0.0 \le |\rho| \le 0.3$ as a low correlation.

1) Temporal Correlation Analysis

If we want to find whether the sampling data of a sensor have a temporal relationship, we should choose a training dataset for the analysis. Assuming that $\{V_{ki}, V_{k(i+1)}, V_{k(i+2)}, ..., V_{k(j-1)}, V_{kj}\}$ is the training dataset of sensor node a_k, we use V_{kj}, $V_{k(j-1)}$, ..., $V_{(j-h)}$ to denote the sampling value of a_k at time T_h. Thus, we obtain the sub-dataset at $T_{(h-1)}$, $T_{(h-2)}$, $T_{(h-3)}$, ..., T_1, as shown in Table 1.

Therefore, we can use Eq. (12) to analyze the relationship of the dataset of T_h and the dataset of another time ($T_{(h-1)}$, $T_{(h-2)}$, ..., T_1). Now, we can define the temporal correlation as follows:

Definition 1: In a training dataset, if the sub-dataset of T_h is highly relevant to one or more sub-datasets of another time ($0.5 < |\rho| \le 1$), the dataset of the sensor node has a high temporal correlation. If the sub-dataset of T_h is only moderately relevant to one or more sub-datasets of another

time ($0.3 < |\rho| \le 0.5$), the dataset of the sensor node has a medium temporal correlation.

2) Spatial Correlation Analysis

If we want to determine whether the sampling data of a sensor have a spatial relationship, we should also choose a training dataset for the analysis. Assuming that $a_{(k+1)}, a_{(k+2)}, ..., a_{(k+i)}$ are the nearest nodes from a_k , we obtain the values listed in Table 2.

Definition 2: In a training dataset, if the sub-dataset of a_k is highly relevant to one or more other sensor node sub-datasets ($0.5 < |\rho| \le 1$), the dataset of the sensor node has a high spatial correlation. If the sub-dataset of T_h is only moderately relevant to one or more other sensor node sub-datasets ($0.3 < |\rho| \le 0.5$), the dataset of the sensor node has a medium spatial correlation.

F. Process of STCAM Decision

STCAM uses the following four algorithms: TCA-LI, TCA-MR, SCA, and STCA. If a sensor node has a tight temporal correlation and does not miss continuous data, STCAM will use the TCA-LI or TCA-MR algorithm to estimate the missing sensor data. Further, if a sensor node has a high spatial correlation, STCAM will use SCA to estimate the missing data. Otherwise, STCAM will choose the STCA algorithm. The process of STCAM decision making is shown in Fig. 2. From these figures, we can conclude the applicable conditions of the four algorithms.

TCA-LI: The training dataset has a high temporal correlation when the type of missing sensor data is single. In contrast, the training dataset has a medium temporal correlation when the training dataset has a low spatial correlation and the type of missing sensor data is single.

TCA-MR: The training dataset has a high temporal correlation, and the type of missing sensor data is continuous. The training dataset has a medium temporal correlation when the training dataset has a low spatial correlation and the type of missing sensor data is continuous.

SCA: The training dataset has a medium temporal correlation when the training dataset has a high spatial correlation. The training dataset has a low temporal correlation when the training dataset has a low spatial correlation, and the training dataset has a low temporal correlation when the training dataset has a medium spatial correlation.

STCA: The training dataset has a medium temporal correlation when the training dataset has a medium spatial correlation.

If the training dataset has a low spatial and temporal correlation, there is no matching algorithm for the estimation.

Table 1. Dataset at time T_h, $T_{(h-1)}$, $T_{(h-2)}$, ..., T_1

Time	Value 0	Value 1	Value 2	...	Value k
T_h	V_{kj}	$V_{k(j-1)}$	$V_{k(j-2)}$	...	$V_{k(j-k)}$
$T_{(h-1)}$	$V_{k(j-1)}$	$V_{k(j-2)}$	$V_{k(j-3)}$	...	$V_{k(j-k-1)}$
$T_{(h-2)}$	$V_{k(j-2)}$	$V_{k(j-3)}$	$V_{k(j-4)}$	...	$V_{k(j-k-2)}$
...	...	...	...	...	...
T_1	$V_{k(j-h)}$	$V_{k(j-h-1)}$	$V_{k(j-h-2)}$	...	$V_{k(j-k-h)}$

Table 2. Dataset of $a_k, a_{(k+1)}, a_{(k+2)}, ..., a_{(k+i)}$ at different times

Node	Value T_i	Value $T_{(i+1)}$	Value $T_{(i+2)}$	...	Value $T_{(i+h)}$
a_k	V_{ki}	$V_{k(i+1)}$	$V_{k(i+2)}$	...	$V_{k(i+h)}$
$a_{(k+1)}$	$V_{(k+1)i}$	$V_{(k+1)(i+1)}$	$V_{(k+1)(i+2)}$	...	$V_{(k+1)(i+h)}$
$a_{(k+2)}$	$V_{(k+2)i}$	$V_{(k+2)(i+1)}$	$V_{(k+2)(i+2)}$	...	$V_{(k+2)(i+h)}$
...	...	...	...	...	...
$a_{(k+i)}$	$V_{(k+i)i}$	$V_{(k+i)(i+1)}$	$V_{(k+i)(i+2)}$	...	$V_{(k+i)(i+h)}$

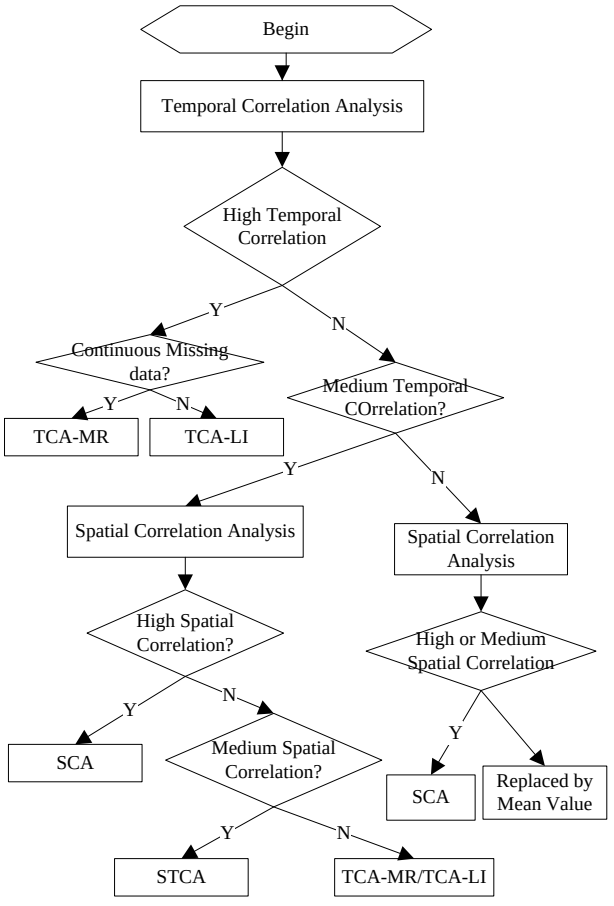

Fig. 2. Process of STCAM decision. STCAM: a model based on spatial-temporal correlation analysis, TCA: temporal correlation analysis, TCA-MR: multiple regression algorithm of TCA, TCA-LI: linear interpolation algorithm of TCA, SCA: spatial correlation analysis, STCA: spatial-temporal correlation analysis.

IV. SIMULATION EXPERIMENTS

The estimation model proposed in this paper is simulated using Java and evaluated over the Intel lab dataset [8] and a traffic dataset of a city in China. The Intel lab dataset is a trace of readings from 54 sensor nodes deployed in the Intel Research Berkeley lab. These sensor nodes collected the light, humidity, temperature, and voltage readings once every 30 seconds. The traffic dataset is a trace of readings from 596 sensor nodes that are deployed on different roads.

To evaluate the accuracy and performance of STCAM, we choose DESM [9] for a comparison. The DESM algorithm is also an estimation approach based on the spatial-temporal correlation, and the result formula is as follows:

$$V_{ki}^{'} = (1-\beta)V_{k(i-1)} + \beta V_{zi}, \quad (13)$$

where $V_{k(i-1)}$ denotes the value of sensor node a_k at (i - 1) time, V_{zi} represents the value of a_z at time i, β denotes the weight of V_{zi}, and DESM chooses a_z as the nearest node from a_k (for a detailed description of DESM, refer to [9]).

To evaluate the four abovementioned algorithms, we need to choose different datasets for testing.

1) Comparison between TCA-LI/TCA-MR and DESM

By analyzing the temporal correlation, we know that the temperature dataset has a high temporal correlation; therefore, we use the temperature dataset of sensor 23 to test the accuracy and performance of TCA-LI/TCA-MR. Firstly, we assume that the 121th, 131th, 141th, ..., 311th data elements are missed; therefore, under this condition, STCAM chooses the TCA-LI algorithm to estimate the missing sensor data. The experiment results are presented in Fig. 3 and Table 3.

We assume that data 121–140 are missing; therefore, under this condition, STCAM chooses the TCA-MR algorithm to estimate the missing sensor data. The experiment results are presented in Fig. 4 and Table 4.

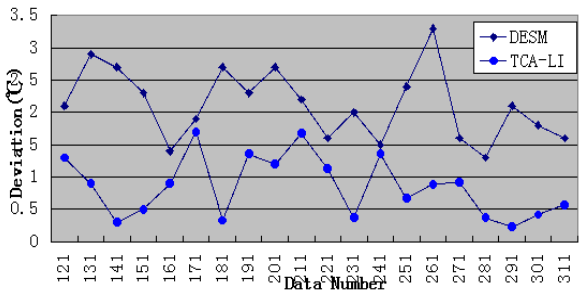

Fig. 3. Comparison of experimental results of TCA-LI and DESM. TCA-LI: linear interpolation algorithm of temporal correlation analysis, DESM: data estimation using statistical model.

Table 3. Performance comparison of TCA-LI and DESM

Index	TCA-LI	DESM	Improvement
Modeling time (sec)	0.0009	0.0013	30.76%
Estimation time (sec)	0.0007	0.0007	0%
Total time (sec)	0.0016	0.0020	20.00%

TCA-LI: linear interpolation algorithm of temporal correlation analysis, DESM: data estimation using statistical model.

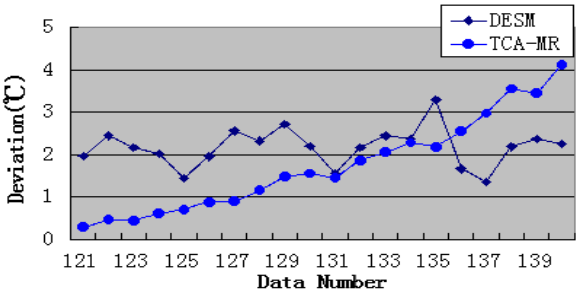

Fig. 4. Comparison of experimental results of TCA-MR and DESM. TCA-LI: linear interpolation algorithm of temporal correlation analysis, DESM: data estimation using statistical model.

Table 4. Performance comparison of TCA-MR and DESM

Index	TCA-MR	DESM	Improvement
Modeling time (s)	0.0109	0.0013	-738.46%
Estimation time (s)	0.0106	0.0007	-1414.28%
Total time (s)	0.0216	0.0020	-980.00%

TCA-LI: multiple regression algorithm of temporal correlation analysis, DESM: data estimation using statistical model.

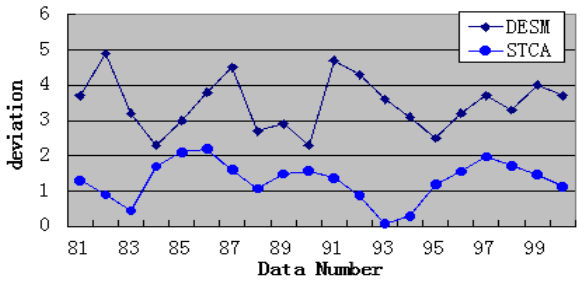

Fig. 5. Comparison of experimental results of STCA and DESM. STCA: spatial-temporal correlation analysis, DESM: data estimation using statistical model.

Table 5. Performance comparison of STCA and DESM

Index	STCA	DESM	Improvement
Modeling time (s)	0.0149	0.0013	-1046.15%
Estimation time (s)	0.0201	0.0007	-2771.42%
Total time (s)	0.0350	0.0020	-1650.00%

STCA: spatial-temporal correlation analysis, DESM: data estimation using statistical model.

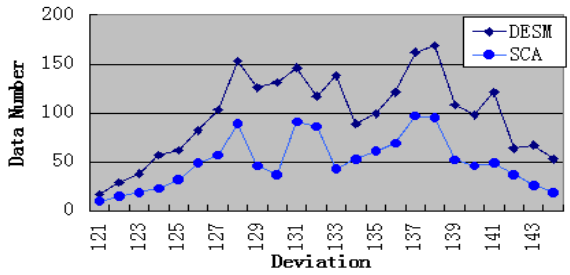

Fig. 6. Comparison of experimental results of SCA and DESM. SCA: spatial correlation analysis, DESM: data estimation using statistical model.

Table 6. Performance comparison of SCA and DESM

Index	SCA	DESM	Improvement
Modeling time (s)	0.0990	0.0013	-507.69%
Estimation time (s)	0.0102	0.0011	-827.27%
Total time (s)	0.0201	0.0024	-737.50%

SCA: spatial correlation analysis, DESM: data estimation using statistical model.

According to Fig. 4, the accuracy of TCA-MR decreases with an increase in the amount of continuous missing sensor data. Therefore, there is a threshold. If the amount of missing sensor data is less than the threshold, TCA-MR exhibits good accuracy, and if the amount of missing sensor data is more than the threshold, TCA-MR is not suitable for estimating the missing sensor data with a high temporal correlation. According to Table 4, the performance of TCA-MR is significantly higher than that of DESM.

2) Comparison between STCA and DESM

By analyzing the temporal and spatial correlation, we know that the humidity dataset has a medium temporal and spatial correlation. Under this condition, STCAM chooses the STCA algorithm to estimate the missing sensor data. We choose the humidity dataset of sensor 11 to test the accuracy and performance of STCA. Assuming that data 81–105 are missing, we obtain the experimental results shown in Fig. 5 and Table 5.

According to Fig. 5 and Table 5, STCA exhibits better accuracy, but the performance is lower than that of DESM.

3) Comparison between SCA and DESM

By analyzing the temporal and spatial correlation, we find that the traffic dataset has a low temporal correlation and a high spatial correlation. Therefore, under this condition, STCAM chooses the SCA algorithm to estimate the missing sensor data. We suppose that data 121–144 of sensor a_6 are missing. The experimental results are shown in Fig. 6 and Table 6.

According to Fig. 6 and Table 6, SCA exhibits better accuracy, but the performance is lower than that of DESM.

V. CONCLUSION

In this paper, we propose a data estimation technique called STCAM, which can discover the correlation of the training dataset, and depending on this correlation and the type of missing sensor data, STCAM can choose one of the most suitable algorithms from SCA-LI, SCA-MR, TCA, and STCA to estimate the missing sensor data. From the simulation result, we conclude that STCAM exhibits good accuracy for the missing sensor data, but in terms of performance, STCAM has a relatively low computational efficiency. Therefore, STCAM can only be deployed at the sink node or in the central server. Moreover, by the simulation, we found that the accuracy of TCA-MR decreases with an increase in the amount of continuous missing sensor data, and this may influence the total accuracy of STCAM, but in the paper, we do not provide an effective solution for this issue. Therefore, in the future, we will conduct further research to fill the gap.

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. NRF-2011-0023076). Further, it was supported by the BB21 project of Busan Metropolitan City.

REFERENCES

[1] L. Pan, H. Gao, H. Gao, and Y. Liu, "A spatial correlation based adaptive missing data estimation algorithm in wireless sensor networks," *International Journal of Wireless Information Networks*, vol. 21, no. 4, pp. 280-289, 2014.

[2] K. Niu, F. Zhao, and X. Qiao, "A missing data imputation algorithm in wireless sensor network based on minimized similarity distortion," in *Proceedings of the 6th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, pp. 235-238, 2013.

[3] S. Ramakrishnan, "Sensing the world," *Jasubhai Digital Media*, vol. 10, no. 1, pp. 26-28, 2003.

[4] M. Halatchev and L. Gruenwald, "Estimating missing values in related sensor data streams," in *Proceedings of the 11th International Conference on Management of Data (COMAD)*, Goa, India, pp. 83-94, 2005.

[5] L. Gruenwald, H. Chok, and M. Aboukhamis, "Using data mining to estimate missing sensor data," in *Proceedings of 7th IEEE International Conference on Data Mining Workshops*, Omaha, NE, pp. 207-212, 2007.

[6] B. S. Yarman, A. Kilinc, and A. Aksen, "Immitance data modelling via linear interpolation techniques: a classical circuit theory approach," *International Journal of Circuit Theory and Applications*, vol. 32, no. 6, pp. 537-563, 2004.

[7] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, pp. 1391-1445, 2009.

[8] S. Madden, Intel lab data [Internet], Available: http://db.csail.mit.edu/labdata/labdata.html.

[9] Y. Li, C. Ai, W. P. Deshmukh, and Y. Wu, "Data estimation in sensor networks using physical and statistical methodologies," in *Proceedings of 28th International Conference on Distributed Computing Systems (ICDCS'08)*, Beijing, China, pp. 538-545, 2008.

Ren Xiaojun
received his B.S. and M.S. degrees from Shandong University of Science and Technology in 2007 and 2010, respectively. Since 2013, he is pursuing his Ph.D. in Data Mining from Dongseo University, Korea. His research interests include artificial intelligence, machine learning, database technology, statistics, and sensor networks.

Hyontai Sug
received his B.S. in Computer Science and Statistics from Busan National University, Korea, in 1983; his M.S. in Applied Computer Science from Hankuk University of Foreign Studies, Korea, in 1986, and his Ph.D. in Computer and Information Science and Engineering from University of Florida, USA, in 1998. He was a researcher of Agency for Defense Development, Korea, from 1986 to 1992 and a full-time lecturer at Pusan University of Foreign Studies, Korea, from 1999 to 2001. Currently, he is a professor at Dongseo University, Korea, from 2001. His research interests include data mining and database applications.

Hoon Jae Lee
received his B.S., M.S., and Ph.D. in Electrical Engineering from Kyungpook National University in 1985, 1987, and 1998, respectively. He was engaged in the research on cryptography and network security at Agency for Defense Development from 1987 to 1998. In 2002, he joined Department of Computer Engineering of Dongseo University as an associate professor, and now, he is a full professor here. His current research interests include security communication systems, side-channel attacks, USN, and RFID security. He is a member of the Korea Institute of Information Security and Cryptology, IEEE Computer Society, IEEE Information Theory Society, etc.