Predicting Missing Values in Wireless Sensor Network using Spatial-Temporal Correlation

Rajeev Kumar, Deeksha Chaurasia, Naveen Chuahan, Narottam Chand
Department of Computer Science & Engineering
National Institute of Technology Hamirpur (INDIA)
rajeev@nith.ac.in, deeksha.nith2015@gmail.com, naveen@nith.ac.in, nar@nith.ac.in

Abstract— **In Wireless Sensor Network (WSN), due to various factors, such as limited power, transmission capabilities of sensors, hardware failures, and network issues like packet collision, unreliable link, and unexpected damage, sensed value does not reach at cluster head/base station. Therefore, data loss in wireless sensor networks is very common. Loss in sensed data greatly reduces the accuracy and analysis of WSN. The data captured by sensor nodes is spatially and temporally correlated due to deployment topology of wireless sensor network. To address problem of missing data, we propose *Prediction using Spatial-Temporal Correlation* (PSTC) algorithm which predicts the missing value of sensor nodes based on spatial and temporal correlation of sensed data.**

Keywords- Wireless sensor network; Missing data; Prediction; Spatial-temporal correlation;

I. INTRODUCTION

Wireless sensor network is the network in which the sensor nodes sense data and send this data to the base station. The combination of sensing and wireless communication has led to the development of wireless sensor network. WSN consists of number of sensor nodes with low processing and limited capabilities like limited storage capacity, battery backup, etc. In wireless sensor network, all the data collected by the sensor nodes is forwarded to the sink node [1]. Wireless Sensor Network have been proposed for various applications including environmental monitoring [2], fire detection [3], object tracking [4], environment reconstruction in cyber space [5], etc. The missing data causes many difficulties in various applications such as accurate environment construction [6], etc. If the missing data/value cannot be predicted accurately, it is very difficult to analyses the overall situation for which sensed data have been collected. At the same time, it is not appropriate to ignore the missing value because it will reduce the accuracy and affects the analysis of the data acquired from sensor nodes. If we simply re-query the missing data it will increase the communication cost, overhead, time and decreases resource efficiency. Data mining can be exploited for extracting/predicting the missing value in WSN, however, we cannot use traditional data mining techniques to find the missing values in wireless sensor network, because of the distributed nature of WSN [7]. Missing value prediction method should be in accordance with the application of wireless sensor network; for example, there exists some real-time application [8] whose deadlines are near to expiry and accuracy in the result is very important.

Sensor network consists of small, low cost sensors that collect and disseminate environmental data. The data from different sensors collectively participate in sending the sensed data to the destination. There are basically two approaches of flow of data from nodes to sink. In first approach, sensor nodes sense the environment and individually every sensor node sends their data directly to the sink. In this approach, the nodes which are far away from sink lose their energy very soon. In second approach, the sensor nodes form the clusters and in each cluster, there is one sensor node that act a cluster head. Every sensor node sends its data to its cluster head.

Every sensor node sends its data to the cluster head using TDMA schedule. Cluster head allots the time to every sensor node for transmitting data. With its respective time slot, every sensor node sends the information to cluster head. If the cluster head did not receive the data within the allotted time slot means the data of that sensor node has been lost. The data at every sensor node is valuable for capturing the complete phenomenon. Now using given prediction technique, the missing sensor data of that sensor node is predicted with accuracy. This technique can accurately predict the missing data by using spatial-temporal correlation.

The contributions of this research paper are:

- We mine sensor datasets at CH and exploit spatial and temporal correlations among the sensor nodes.
- Based on the observed correlations, we design the PSTC algorithm to accurately predict the missing data at CH/sink.
- We perform simulation to evaluate the performance of proposed algorithm PSTC.

The rest of the paper is organized as follows. Section II discusses about related work. Proposed prediction algorithm PSTC is discussed in Section III. In Section IV, we have given performance evaluation of our proposed algorithm. Section V gives the conclusion.

II. RELATED WORK

Le Gruenwald et al. proposed WARM (Window Association Rule Mining) [9] to deal with missing data, this technique uses 2-frequent item sets association rule mining which means it can discover the relationships only between two sensors and ignore the cases where missing values are related with multiple sensors. SPIRIT (Streaming Pattern dIscovery in multIple Time series) [10] technique is proposed by Spiros Papadimitriou et al. in which they have used auto regression for finding correlation using hidden variables inside the history data of a sensor. It estimates missing data by predicting changes in data patterns using hidden variables without buffering of stream values. Nan Jiang et al. proposed CARM (Closed frequent item-set Association Rule Mining) [11] technique which can derive the most recent association rules between sensors based on the current closed item sets in the current sliding window. The closed item set mining effectively imputes missing values as well as achieve time and space efficiency. Le Gruenwald et al. proposed data estimation technique FARM (Freshness Association Rule Mining) [12], to estimate value for missing sensors. Authors used the association rules by considering the freshness of data and it implemented a data compaction scheme to store history data. Yingshu Li et al. proposed DESM (Data Estimation using Statistical Model) [13], that estimates the values using the spatial and temporal correlation of the nodes. In this model, the whole network is divided into cells using a grid. This model prolongs the network lifetime with high accuracy. [14] presented by Yuan Li et al., is a spatial-temporal imputation technique for estimating missing data in wireless sensor network, if the environment is highly correlated in time and space.

Zaifie Liao et al. [15] proposed the Fuzzy K-means Clustering Algorithm over sliding window to reduce the impact of data volume and provide a better tool when the clusters are not well separated as it use sliding window and fuzzy set. Le Gruenwald et al. proposed DEMS (Data Estimation for Mobile Sensors) [16], that mines spatial and temporal relationship among mobile sensors with the help of virtual static sensor. It divides the entire monitoring area into hexagons based on user defined radius. Each hexagon corresponds to a virtual static sensor, trajectory pattern tree is used to predict the mobile sensors location as well as reading. Liqiang Pan et al. proposed AKE (Applying K- nearest neighbor Estimation) [17], the algorithm is based on the spatial correlation more than the temporal correlation, and estimates the missing data utilizing multiple neighbor nodes jointly rather than independently, so that a stable and reliable estimation performance can be achieved. Sneha Arjun Dhargalkar et al. [18] proposed the method to estimate the missing value present in the existing databases by taking the approach WARM and AKE so that the data is always complete and accurate. For the similar value of the other sensor nodes impute the average value of missing sensor node in a given window size. Liqiang Pan et al. [19] proposed the estimation algorithm based on the spatial correlation of sensor data. This technique estimates the data using multiple neighbor nodes jointly. The estimation equation can be adjusted adaptively for different missing data.

III. Prediction Using Spatial-Temporal Correlation (PSTC)

Since missing data is inevitable in wireless sensor network applications and causes many problems, development of missing data prediction algorithms is essentially required. In order to reduce the problem/difficulties erupted because of missing data we have proposed Prediction using Spatial-Temporal Correlation (PSTC) to estimate the missing/lost data in WSN. Proposed algorithm uses spatially and temporally correlated sensor nodes to predict the missing value. After deployment of the sensor nodes, cluster formation takes place, which results into clusters of those sensor nodes which are temporally and spatially correlated. After cluster formation sensor nodes sense data and send it to cluster head. Cluster head collects the sensed data, processes it and sends it to sink. If cluster head identifies any data loss from any of its members, then cluster head uses PSTC to predict the missing value. Prediction of missing value at cluster head has several advantages as compared to requesting data from the sensor node. Figure 1 shows the framework of the proposed algorithm.

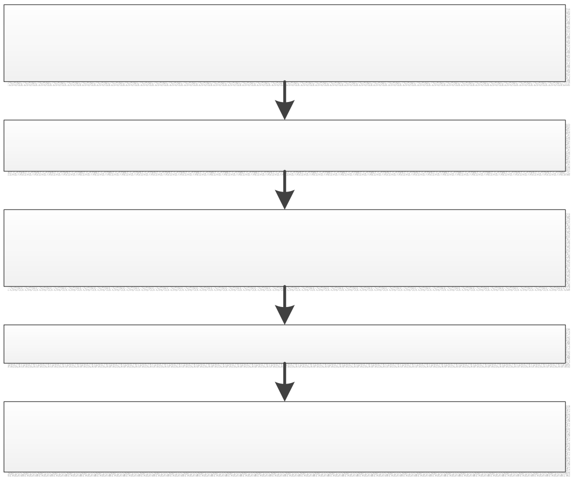
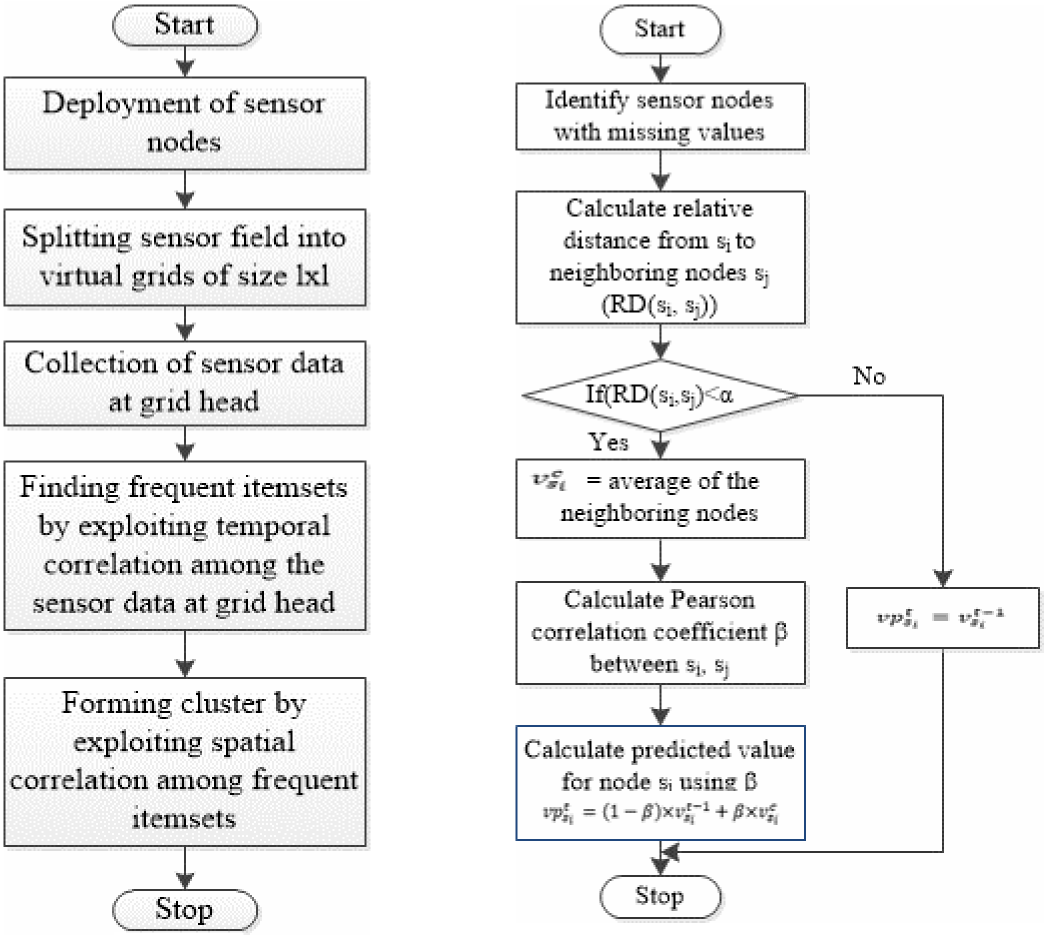


Figure 1: Framework of proposed prediction algorithm.

A. Formation of Cluster

In this section, we describe an adaptive clustering protocol for wireless sensor networks which we have proposed in [20]. To exploit spatial correlation, initially we divide the given sensor field into virtual grid like clusters. Thereafter, data mining techniques are applied to exploit the temporal correlation among sensors based upon their sensed values over a time interval. The initial selection of cluster head (CH) within a grid/cluster is based upon the choice of a node that is located nearest to the centroid of the grid. Initially CH collects data of its members and generates frequent itemsets exploiting temporal correlation among its member sensor nodes. These frequent itemsets further processed by sink to form temporally and spatially cohesive clusters. Figure 2(a) gives the steps involved in the cluster formation.

If within specified time interval the sensed data is not received at CH, then CH assumes that there is loss of data due to any of the reason. Once data loss identified by CH, then CH uses proposed PSTC algorithm to predict the missing value of respective cluster member. Figure 2(b) gives the steps involved in the prediction algorithm.


Start
Deployment of sensor nodes
Splitting sensor field into virtual grids of size lxl
Collection of sensor data at grid head
Finding frequent itemsets by exploiting temporal correlation among the sensor data at grid head
Forming cluster by exploiting spatial correlation among frequent itemsets
Stop
Start
Identify sensor nodes with missing values
Calculate relative distance from s_i to neighboring nodes s_j (RD(s_i, s_j))
If(RD(s_i,s_j)<α
No
Yes
$v^c_{s_i}$ = average of the neighboring nodes
Calculate Pearson correlation coefficient β between s_i, s_j
$vp^t_{s_i} = v^{t-1}_{s_i}$
Calculate predicted value for node s_i using β
$vp^t_{s_i} = (1-\beta)\times v^{t-1}_{s_i} + \beta\times v^c_{s_i}$
Stop

(a)

(b)

Figure 2: (a) Steps involved in cluster formation (b) steps involved in prediction algorithm.

Prediction Algorithm

After identification of loss of data cluster head runs the prediction algorithm. As the algorithm uses spatial-temporal correlation, so first CH calculates distance from missing sensor node s_x to other neighboring sensor nodes. If the distance between s_x to other sensor nodes is not less than a predefined threshold distance (α) value then CH assumes average value of spatially correlated neighbors of s_x as its current value. CH further process collected data and sends to sink. If the distance between s_x to other neighboring nodes is less than threshold distance (α) value then PSTC algorithm uses Pearson correlation coefficient to exploit spatial correlation among such nodes. Pearson correlation gives the linear relationship between two sensor nodes. Pearson correlation between two sensor nodes is given by (1).

$$\rho_{s_x s_y} = \frac{Cov(s_x, s_y)}{\rho_{s_x}\rho_{s_y}} \qquad (1)$$

Covariance between two sensor nodes is given by (2).

$$Cov(s_x, s_y) = \frac{\sum_{i=1}^{t}(v^i_{s_x} - \overline{v_{s_x}})(v^i_{s_y} - \overline{v_{s_y}})}{t-1} \qquad (2)$$

Standard deviation of two sensor nodes can be computed at cluster head as CH maintains the history data of its members. Standard deviation between two sensor nodes is given by (3) and (4).

$$\rho_{s_x} = \sqrt{\frac{\sum_{i=1}^{t}(v^i_{s_x} - \overline{v_{s_x}})^2}{t-1}} \qquad (3)$$

$$\rho_{s_y} = \sqrt{\frac{\sum_{i=1}^{t}(v^i_{s_y} - \overline{v_{s_y}})^2}{t-1}} \qquad (4)$$

By substituting (2), (3) and (4) into (1), Pearson correlation between two sensor nodes is given by (6).

$$\rho_{s_x s_y} = \frac{\sum_{i=1}^{t}(v^i_{s_x} - \overline{v_{s_x}})\,(v^i_{s_y} - \overline{v_{s_y}})}{\sqrt{\sum_{i=1}^{t}(v^i_{s_x} - \overline{v_{s_x}})^2}\,\sqrt{\sum_{i=1}^{t}(v^i_{s_y} - \overline{v_{s_y}})^2}} \qquad (5)$$

A correlation of -1 means there is a perfect negative association between variables, 0 means there is no linear relationship between the two variables and +1 means that there is perfect positive association between variables. Algorithm 1 gives the steps of PSTC.

Table 1: Notation used

Symbol	Meaning
$vp_{s_i}^t$	Predicted value for node s_i at time t
$v_{s_x}^c$	Average value of the neighboring nodes of node s_x at time t
α	User specified threshold distance
β	Pearson's linear correlation coefficient among sensor node s_i and s_j
RD(s_i, s_j)	Relative distance between node s_i and s_j

Algorithm 1: Prediction using Spatial-Temporal Correlation (PSTC)

```
For each sensor node $s_x$, whose value is missing at CH
Calculate RD($s_x, s_j$) for all $s_j$, such that $s_j$ belongs to CH and $s_j$ is neighbor of $s_x$
For all nodes $s_y$ such that RD($s_x, s_y$) $< \alpha$
    $v_{s_x}^c = \frac{\sum_{i=1}^{n} v_{s_{yi}}^t}{n}$
Compute $\beta = \frac{\sum_{i=1}^{t}(v_{s_x}^i - \overline{v_{s_x}})(v_{s_y}^i - \overline{v_{s_y}})}{\sqrt{\sum_{i=1}^{t}(v_{s_x}^i - \overline{v_{s_x}})^2}\sqrt{\sum_{i=1}^{t}(v_{s_y}^i - \overline{v_{s_y}})^2}}$
$vp_{s_x}^t = (1-\beta) \times v_{s_x}^{t-1} + \beta \times v_{s_x}^c$
if there is no such node $s_y$ such that RD($s_x, s_y$) $< \alpha$
    $vp_{s_x}^t = \frac{\sum_{i=1}^{t-1} v_{s_x}^i}{t-1}$
CH uses $vp_{s_x}^t$ and aggregates collected data and sends to Sink
```

IV. Performance Evaluation

In this section, we evaluate the performance of proposed algorithm Prediction using Spatial-Temporal Correlation (PSTC) Through simulating in MATLAB. We first define simulation parameters, and performance metric used. Simulation results are compared with LIN [19] and KNN [17] approach. Linear (LIN) method is temporal correlation based method to estimate the missing values. This approach is based on linear interpolation model. KNN method is naive spatial correlation method to estimate the missing values. The estimation of KNN is done by weighted average of all neighboring nodes value.

A. Simulation Parameters

We consider a flat and square two-dimensional sensor field of size 200m×200m in which sensor nodes are randomly deployed. All nodes are homogeneous. Various simulation parameters are listed in Table 2.

B. Performance Metrics

To check the performance of proposed algorithm Root Mean Square Error (RMSE) is used as performance matric. RMSE is the square root of the mean/average of the square of the error, which is magnitude not the percentage. RMSE is used for regression analysis and to verify results (to compare different models or accuracy of prediction). The formula of RMSE can be expressed as:

$$RMSE = \sqrt{\frac{(vp_{s_i}^t - v_{s_i}^t)^2}{n}} \qquad (6)$$

Parameter	**Default Value**	**Range**
Network size (side of square sensor field)	200m	50m ~ 300m
Number of nodes	400	100~500
Transmission range (R)	100m	40m ~ 140m
Initial energy of node	2 Joule	
Sink location	(0, 0)	
Data packet size	64 KB	
E_{elec}	50 nJ/bit	
$_{fs}$	10 pJ/bit/m^2	
$_{mp}$	0.00134 pJ/bit/m^4	
Data rate	512 Kbps	
α	User specified spatial factor	60m

C. Results and Discussion

- **Effect of number of missing values on RMSE**

From the Figure 3, it is observed that when the number of missing values in a cluster is small, error in prediction in missing value is also small. Error in prediction of missing values increases with the number of missing data increase. Results of PSTC are better than LIN and KNN. In LIN, linear interpolation model is used to predict the missing value and in KNN nearest neighbor's data is used to predict the missing value of the sensor node. PSTC exploits temporal as well as spatial correlation among sensor nodes, due to which error in prediction is less.

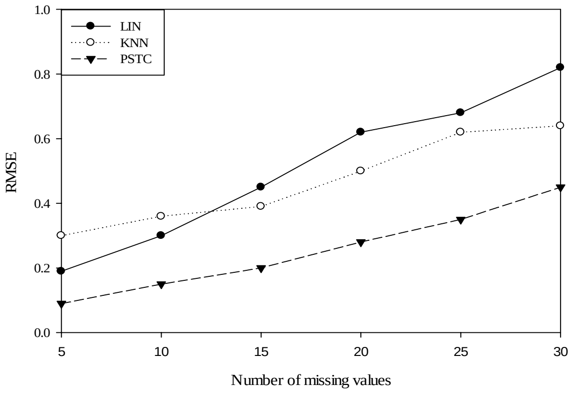


Figure 3: Effect of number of missing values on RMSE

- **Effect of number of neighboring nodes on RMSE**

From the Figure 4, it is observed that when the number of neighboring nodes to a node with missing value is less, then error in prediction of missing value is more. Error in prediction of missing values decreases with the increase in number of neighboring nodes. PSTC performs better than LIN and KNN. In case of PSTC, as number of neighboring nodes increases, more number of nodes are spatially and temporally cohesive as more number of nodes are located under the distance < α, which helps in predicting missing value of the given sensor node. In case of LIN method RMSE remains almost constant as increase in number of neighboring nodes does not improve in predicting missing value. In KNN method RMSE decreases as number of neighboring nodes increases.

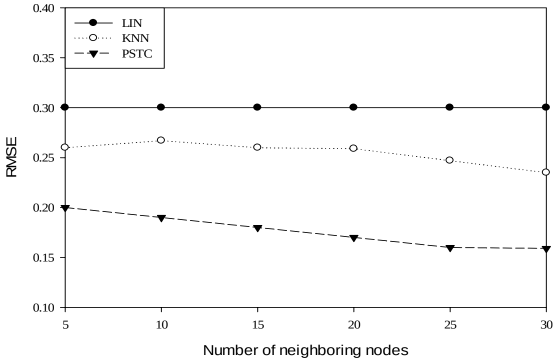


Figure 4: Effect of number of neighboring nodes on RMSE

- **Effect of sampling interval on RMSE**

Figure 5 shows the result of RMSE as a function of sampling interval of sensor nodes. From the graph, it has been observed that when the sampling interval is small, error in predicting the missing value is small. As if sampling interval is small then there is temporal correlation among the values which results in less error in predicting missing value. When sampling interval increases then there will be decrease in temporal correlation among the sensed values which leads to increase in error. In LIN, increase in sampling interval leads to more error in predicting the missing value of the sensor node due to loss in temporal correlation among the sensor nodes. Our approach PSTC outperforms LIN as well as KNN. This is because of the temporal and spatially cohesive nature of the cluster formed, which helps in better prediction of missing value.

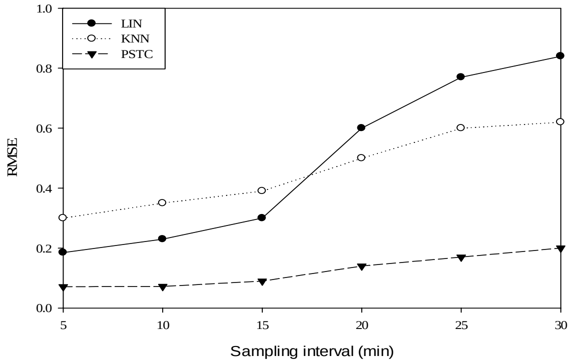


Figure 5: Effect of sampling interval (min) on RMSE

V. Conclusion

In this paper, we present a technique for predicting missing values of sensor node as well as to maintain and improve the efficiency of the wireless sensor network. Lost sensor data is unavoidable in wireless sensor network, and it causes many complications in its various applications. PSTC algorithm predicts the missing value by using spatial-temporal correlation among the sensor nodes. Proposed algorithm PSTC aims to predict the missing data with less error. It has been established that PSTC algorithm also performs better than other algorithms.

References

[1] E. Ben Hamida and G. Chelius, "Strategies for data dissemination to mobile sinks in wireless sensor networks," *IEEE Wireless Communications*, vol. 15, no. 6, 2008, pp. 31–37.

[2] D. Kumar, "Monitoring Forest Cover Changes Using Remote Sensing and GIS: A Global Prospective," *Research Journal of Environmental Sciences*, vol. 5, no. 2, 2011, pp. 105–123.

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, 2002, pp. 393–422.

[4] S. H. Chauhdary, A. K. Bashir, S. C. Shah, and M. S. Park, "Eoatr: Energy Efficient Object Tracking by Auto Adjusting Transmission Range in Wireless Sensor Network," *Journal of Applied Sciences*, vol. 9(24), 2009, pp. 4247–4252.

[5] L. Kong, D. Jiang, and M. Y. Wu, "Optimizing the Spatio-Temporal Distribution of Cyber-Physical Systems for Environment," *International Conference on Distributed Computing Systems*, no. 2, 2010, pp. 179–188.

[6] L. Kong *et al.*, "Data Loss and Reconstruction in Wireless Sensor Networks," *IEEE INFOCOM*, 2013, pp. 1654–1662.

[7] A. Mahmood, K. Shi, S. Khatoon, and M. Xiao, "Data Mining Techniques for Wireless Sensor Networks : A Survey," *International Journal of Distributed Sensor Networks*, vol. 2013, 2013, pp. 1–24.

[8] T. Arampatzis, J. Lygeros, and S. Manesis, "A Survey of Applications of Wireless Sensors and Wireless Sensor

Networks," *Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation Intelligent Control, 2005.*, 2005, pp. 719–724.

[9] M. Halatchev and L. Gruenwald, "Estimating Missing Values in Related Sensor Data Streams," *Advances in Data Management*, 2005, pp. 83–94.

[10] S. Papadimitriou and C. Faloutsos, "Streaming Pattern Discovery in Multiple Time- Series Streaming Pattern Discovery in Multiple Time-Series," *International Conference on Very Large Data Bases (VLDB)*, 2005, pp. 697–708.

[11] N. Jiang and L. Gruenwald, "Estimating Missing Data in Data Streams," *Advances in Databases: Concepts, Systems and Applications. DASFAA 2007. Lecture Notes in Computer Science*, vol. 4443, 2007.

[12] L. Gruenwald, H. Chok, and M. Aboukhamis, "Using data mining to estimate missing sensor data," *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2007, pp. 207–212.

[13] Y. Li, C. Ai, W. P. Deshmukh, and Y. Wu, "Data Estimation in Sensor Networks Using Physical and Statistical Methodologies," *2008 The 28th International Conference on Distributed Computing Systems*, 2008, pp. 538–545.

[14] Y. Y. Li and L. E. Parker, "Classification with missing data in a wireless sensor network," *IEEE SoutheastCon 2008*, 2008, pp. 533–538.

[15] Z. Liao, X. Lu, T. Yang, and H. Wang, "Missing Data Imputation: A Fuzzy K-means Clustering Algorithm over Sliding Window," *International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 3, 2009, pp. 133–137.

[16] L. Gruenwald, M. Sadik, R. Shukla, and H. Yang, "DEMS: A Data Mining Based Technique to Handle Missing Data in Mobile Sensor Network Applications," *Proceedings of the Seventh International Workshop on Data Management for Sensor Networks*, 2010, pp. 26–32.

[17] L. Pan and J. Li, "K-Nearest Neighbor Based Missing Data Estimation Algorithm in Wireless Sensor Networks," *Wireless Sensor Network*, vol. 2, no. 2, 2010, pp. 115–122.

[18] S. A. Dhargalkar and A. U. Bapat, "Determining Missing Values in Dimension Incomplete Databases using Spatial-Temporal Correlation Techniques," *Proceedings of 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014*, no. 978, 2014, pp. 601–606.

[19] 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, 2014, pp. 280–289.

[20] R. Kumar, N. Chauhan, and N. Chand, "Adaptive Clustering in Wireless Sensor Networks."