Collective Prediction exploiting Spatio Temporal correlation (CoPeST) for energy efficient wireless sensor networks

Muruganantham ARUNRAJA[1] , Veluchamy MALATHI[2]
[1] Anna University, Regional Centre, Madurai, Tamil Nadu, India
[E-mail: researcharunraja@gmail.com]
[2] Anna University, Regional Centre, Madurai, Tamil Nadu, India
[E-mail: vmeee@autmdu.ac.in]
*Corresponding author: Muruganantham Arunraja

Received September 24, 2014; revised January 16, 2015; accepted June 9, 2015; published July 31, 2015

Abstract

Data redundancy has high impact on Wireless Sensor Network's (WSN) performance and reliability. Spatial and temporal similarity is an inherent property of sensory data. By reducing this spatio-temporal data redundancy, substantial amount of nodal energy and bandwidth can be conserved. Most of the data gathering approaches use either temporal correlation or spatial correlation to minimize data redundancy. In Collective Prediction exploiting Spatio Temporal correlation (CoPeST), we exploit both the spatial and temporal correlation between sensory data. In the proposed work, the spatial redundancy of sensor data is reduced by similarity based sub clustering, where closely correlated sensor nodes are represented by a single representative node. The temporal redundancy is reduced by model based prediction approach, where only a subset of sensor data is transmitted and the rest is predicted. The proposed work reduces substantial amount of energy expensive communication, while maintaining the data within user define error threshold. Being a distributed approach, the proposed work is highly scalable. The work achieves up to 65% data reduction in a periodical data gathering system with an error tolerance of 0.6°C on collected data.

***Keywords*:** wireless sensor network, data reduction, data prediction, similarity based clustering

1. Introduction

With the advancements in MEMS, chip integration and Radio frequency technologies, WSNs are applied to a variety of applications including environmental monitoring [1], military surveillance [2], industrial process controls [3], smart spaces and many more. WSN are intended for distributed long term data gathering, while maintaining required accuracy. In a WSN, each sensor node is a self contained system consists of sensing, computing and communicating elements. A major constraint of sensor nodes is their finite energy source. Wireless communication is a primary energy consuming functionality, where sensing can also play an important role depending on the particular type of sensing performed. On the other hand, the computation is the least energy consuming activity. One of the main objectives in deploying WSN is to achieve a resilient large scale data collection, while maintaining sufficiently high quality of the collected data. The amount of data transmitted from sensor nodes and the number of active sensor nodes at the time instant has high impact on the cost of the distributed monitoring process. In MICAZ mote, the energy costs for transmission and reception of one bit are 600nJ and 670nJ with 3.5nJ computation energy per clock cycle. Whereas in TELOSB mote, the energy costs of transmission, reception and computation are 720nJ, 810nJ and 1.2nJ respectively [4]. This reveals that the communication cost is much higher than computation cost.

In a periodic data collection approach [5], nodes sense the environment and transmit the data of interest continuously over time, only by which a finest data granularity can be obtained. On one hand, several environmental and habitat monitoring applications require periodic long-term data collection, as the gathered data make sense only if the data collection procedure lasts for months or even years continuously without interruption. On the other hand, sensor nodes are often energy constrained and deployed in harsh environments, hence data collection strategy must be energy conscious to prolong the network lifetime as much as possible. Due to the pervasive existence of Spatio-temporal correlation in the sampled data, huge portion is constituted by redundant data [6]. These redundant data do not have any informational value, but consumes the network resources substantially. By effectively modeling and exploiting spatial and temporal correlation, huge amount of data transmission can be reduced. This motivates the need for a comprehensive spatio-temporal redundancy aware data gathering approach that achieves high energy efficiency without much compromising on the accuracy of collected data.

The spatial correlation is high between the physically nearer sensors. If the sensor nodes are closer, similar their observations, hence the observations of a sensor node may be predicted from that of its neighboring sensor nodes with high confidence. The magnitude and trend similarity of data generated by these nodes confirms their close spatial correlation. Within the correlated network, only a subset of sensors may report the data and the rest can be approximated. The higher the sampling frequency of a sensor node, similar the consecutive data, hence the future readings of a sensor node can be predicted based on the recent previous readings from the same node. The temporal correlation can be used to estimate the trend of the signal, through which the future data can be predicted.

In the proposed work, we attempted to exploit both spatial correlation and temporal correlation among the intra sensor and inter sensor data to reduce the communication expense without losing significant accuracy. The work groups the sensors with similar observations

inside a cluster into multiple sub clusters. A sub cluster is represented by its sub cluster head (SCH). The SCH node constructs a temporal correlation model using an adaptive filter based on the recent history of data.The model is communicated to the CH and other sub cluster members (SCMs). Based on the model, prediction takes place at CH, SCH and SCMs at every sampling instant. The SCH updates the trend changes by updating the filter coefficients appropriately. The sub cluster members transmit data only when the difference between the model predicted data and observed data exceeds the user defined threshold. This system can effectively reduce the communication cost of periodic reporting framework, while the user-defined accuracy is guaranteed for all sensor nodes.

The proposed system has been evaluated on a synthetic data set with various correlation degrees. The work shows better efficiency in terms of energy. The accuracy of collected data is also shown to be good. The sub clusters have better balanced the nodal energy without compromising on data accuracy. The rest of the paper is organized as follows: Section 2 briefly discusses the related works. Section 3 elaborates on the temporal and spatial correlation based data estimation. Section 4 evaluates the CoPeST by trace-driven simulations. Section 5 concludes the paper.

2. Related Work

Energy efficient functionality is the key issue in the design of wireless sensor networks. Limited energy being a bottleneck for most of the WSN applications, numerous works have been done on energy conservation in WSN. Anatasi et al [7] have discussed key directions of energy conservation in WSN, where duty cycling, data driven and mobility based approaches are discussed. Data driven approaches aim at reducing redundant data transmissions, thus by conserving energy and preserving data integrity. The reduced data traffic results in nodal energy conservation, bandwidth conservation and data collision avoidance.

The temporal correlation between consecutive data can be exploited to reduce the temporally redundant data through multiple approaches [8]. Data prediction reduces the communication overhead by estimating the temporal correlation and predicting the future data from recent data history. Several data prediction approaches, exploit temporal correlations between sensory data using linear regression methods [9], but suffers from reduced accuracy due to lack of adaptability towards dynamic variations in input signal. In [10], Auto Regressive Integrated Moving Average (ARIMA) based methods is used to predict the sensor data from previous values. ARIMA needs a great number of basic data, thus computationally expensive and is poor at predicting series with turning points. In [11], the prediction is executed using Principal Component Analysis (PCA) that necessitates the prior model definition. In Predicitve Storage architecture for WSN (PRESTO) [12], high tier proxies construct a model that captures correlations in the data observed at each low tier sensor. The remote sensors check the sensed data against this model and push data only when the observed data deviates from the values predicted by the model, thereby capturing anomalous trends. In PRESTO, the data models are reconstructed only periodically and the trend changes are not addressed immediately. In our proposed work, we use a model free prediction filter based on the Least Mean Square (LMS) algorithm [13] for exploiting temporal correlation. The proposed approach is computationally light weight and highly adaptive to the data dynamics.

In [14], an energy efficient data gathering method is discussed, where the fraction of active sensors is regulated based on the spatial characteristics of the observed phenomenon. In [15], a linear model is proposed to capture the spatial correlation in sampling data from different sources. With this model, most sensor nodes can be put into sleeping mode, and their readings

can be estimated with definite accuracy by using the linear combination of data from working sensor nodes. However, in the real world, a lot of systems may not be linear. Furthermore, the method of choosing the right working nodes has not been discussed in [15]. An Adaptive Sampling Approach (ASAP) [16] creates sub clusters with correlated sensor nodes and selects a subset of samplers from sub cluster through which continuously collect the data. The non-sampler data are predicted using spatial correlation. ASAP uses probabilistic models validated only during forced sampling periods, thus does not guarantee the predicted data's error bound. Anomalous trends between forced samples may go unnoticed, which is the common difficulty with most of the spatial data reduction schemes. The proposed work follows a model based push approach, where the spatial correlation is verified on every data collection round and the anomalies are corrected immediately.

There are few data gathering approaches that utilize both spatial and temporal correlations between the sensor data to reduce the communication overhead. In [17], the authors have defined the theoretical framework for exploiting spatio temporal correlation. Here they suggested a correlation based MAC that reduces the number of representative nodes, while preserving measurement distortion at the low. The minimal reporting rate is also recommended to achieve required event detection reliability rate with minimum resource utilization. A tiny-model query system called Barbie-Q (BBQ) [18] uses multivariate Gaussian joint distribution to capture the correlations of sensor readings. It samples a small fraction of sensor data from a WSN and utilizes a Gaussian joint distribution model to estimate the non-sampled sensor readings. However, these kinds of models need an expensive long training phase and a complete data set of every sensor node within a sufficiently long period. Second, the correctness of this kind of models requires a continuous model update which needs periodically gathering the data generated by every sensor node and disseminating the update information to related sensor nodes. An Energy-Efficient Data Collection Framework (EEDC) [19] selects nodes with similar data to form clusters. Within a cluster, one sensor node data can be approximated from other node's data. Thus inside a cluster, sensor nodes are scheduled to work alternatively to save energy. Temporal correlation is exploited by piecewise linear approximation, where the time series is reduced into short line segments. EEDC suffers from scalability issue as being a centralized approach. Secondly EEDC cannot guarantee the data error bound for the un-sampled nodes. In [20], the authors proposed a technique called Self-Based Regression (SBR) technique which works on a tree based aggression network. Here the data stream of a sensor node is divided into smaller data packets of length W. The packet that most closely representing the entire data is selected and communicated. Here the spatial correlation is exploited by periodical snooping of neighboring nodes and the most correlated neighbors are suppressed. Every time node has to send heart beat message to validate the data correlation between its own and the representative data, which consumes a hefty amount of nodal data. It requires large memory and lengthy data computations, which makes the system less energy efficient. This method is comfortable for query based data collection, whereas our method can be applied to a continuous data collection.

The proposed work is designed in such a way to accommodate the goals of an ideal data gathering system. Being a distributed system, this work can be applied to WSN of any magnitude. The system exploits the spatial correlation through the construction of similarity based sub clusters represented by a single node. Temporal correlation is exploited by dual prediction based reporting. Both the approaches are combined into collective prediction to achieve spatio-temporal correlation based data reduction. The system uses light weight algorithms, that are most suitable for resource constrained WSN. The contribution of the proposed system is in two folds. The system alters the conventional dual prediction based

reporting scheme into the collective prediction scheme by considering the spatial correlation between closely located nodes. Unlike other spatial correlation based reporting systems, the proposed system is a highly supervised system that detects and corrects anomalies at the time of the occurrence itself, without much overhead.

3. Collective Dual Prediction

3.1. System Model & Overview

The proposed system follows three layer architecture. The bottom layer consists of N number of nodes randomly distributed over the field of interest. Each node is a self-contained system with its own sensing, computing and communication modules powered by a finite energy source. The second layer consists of node clusters attributed by a group of spatially nearer nodes associated with a high energy cluster head. The nodes transmit their data to the CH, where it is aggregated and forwarded to the base station through a backbone constructed by cluster heads. The third layer is built over the clusters. Here the cluster members are subdivided into numerous sub clusters of closely correlated nodes. The sub clusters are represented by an SCH. The data generated by SCH is deemed to represent the whole sub cluster.

The proposed work attempts to reduce the data communication over the network by exploiting spatial and temporal correlation between the sensor data. The spatial correlation between the sensor data is estimated and the highly correlated sensors are grouped into sub clusters. Each sub cluster is represented by an SCH node, thus the redundant data from neighboring nodes are suppressed. The temporal correlation in the sensor data series is estimated using an LMS filter and is used to predict the future data. Since the data are predicted, only a subset of data, that deviates from the desired data are transmitted. Hence the method completely filters out spatially and temporally redundant data. Here a collective prediction approach is introduced, through which anomalous trends from the non-sampler nodes are also identified and communicated to the sink.

The functionality of the proposed system is in four phases. In the first phase energetic nodes are selected as cluster heads using weight based passive clustering method and clusters are constructed around them. Next, the CH collects data from its members and further divides them into sub clusters based on their data similarity. The node with highest energy (SCH) represents the sub cluster. In the third phase, the SCH constructs a temporal correlation model from its previous observations using an LMS based filter. Through the model, future data can be predicted within a user defined error tolerance. The model is shared with the CH and sub cluster members. In the fourth phase, at every sampling instant, the SCH compares the predicted data with the observed data. When the prediction deviates from the observation beyond the temporal error threshold for certain consecutive rounds, the model is updated.

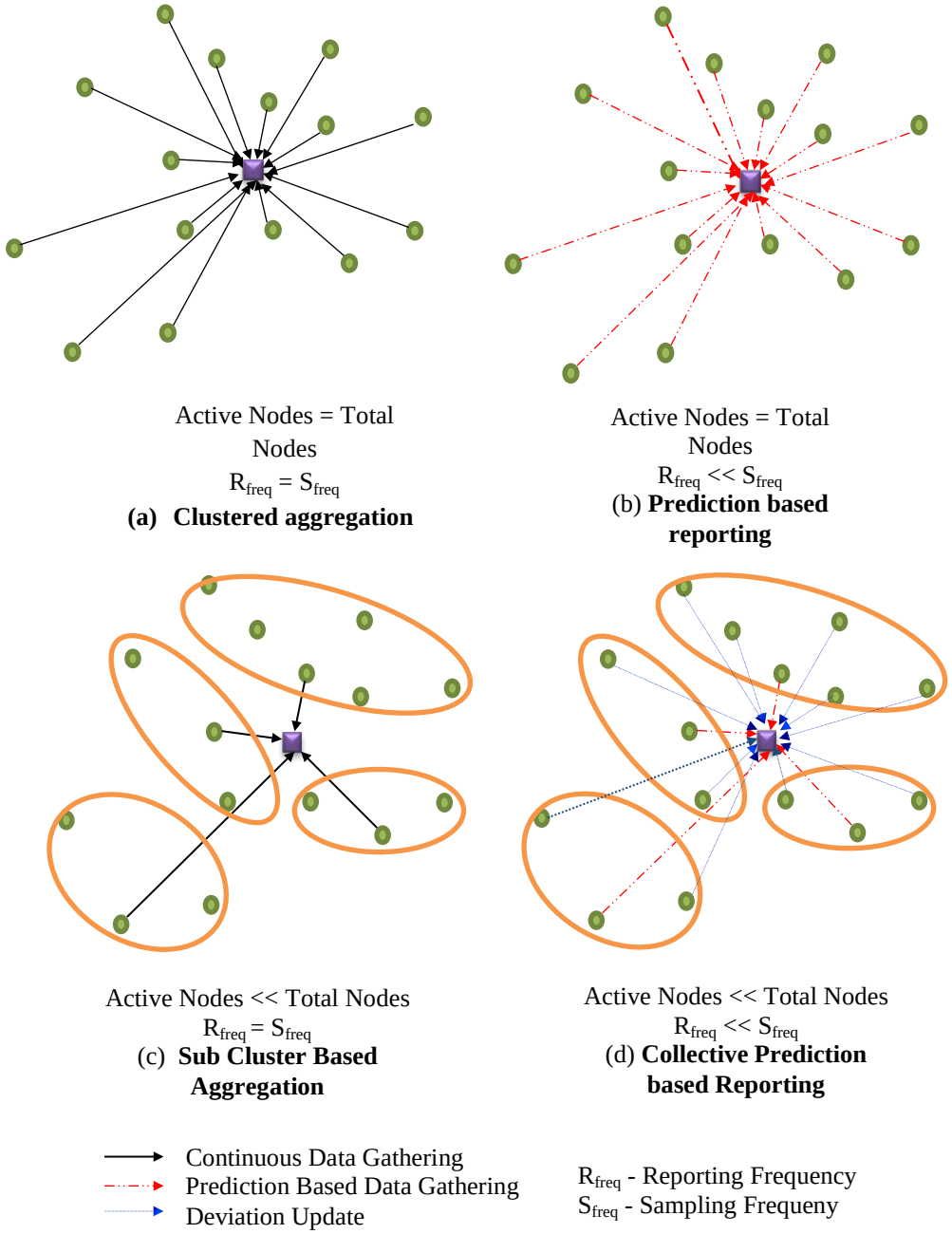

Fig. 1. Evolution of CoPeST Frame work

At every sampling instant, the sub cluster members compare the predicted data with their own observations and communicate the data, when the deviation is larger than the spatial error threshold. At every sampling instant, the CH predicts the data based on the model and checks for the updates from SCH and communications from the sub cluster members. If there is no update from the SCH, the model is considered accurate for the time instant. Then the estimated data are considered as the approximation for the whole sub cluster. If there is a communication

from the sub cluster member, the corresponding data are replaced in that sub cluster member's data series.

Fig. 1 indicates step by step evolution of our proposed collective prediction framework. **Fig. 1. (a)** shows the conventional clustered aggregation scheme, where all the cluster members continuously update their sensed data to the CH. Hence consumes huge amount of residual energy. **Fig. 1. (b)** indicates the prediction based reporting approach, where the nodes report only a subset of the total data sensed, the rest is predicted. Hence reduces the temporally redundant data and improves energy efficiency. **Fig. 1. (c)** shows the sub cluster based aggregation scheme, where only a portion of total nodes report their data to CH. Other nodes closer to the active nodes stay idle. Thus the spatial redundancy of data is reduced and energy is conserved. **Fig. 1. (d)** shows our proposed work combines dual prediction with sub clustering. Hence, both temporally and spatially redundant data are suppressed, to achieve comparatively higher energy efficiency than the previous methods.

3.2. Energy efficient passive clustering

In most of the WSNs the physically nearer nodes are clustered to achieve scalability, bandwidth conservation and routing comforts. The clusters distribute the computational load of the network, by undertaking various complex estimations at the CH itself. Since our algorithm needs to perform numerous computations on different data series, we first divide the network into groups of spatially nearer nodes headed by an energetic node. The proposed work can be applied to most of the existing clustering algorithms, with the following constraints. The members of a cluster should be able communicate with their CH directly, since it is essential for prediction based reporting and anomaly detection. As the variations in spatial correlation necessitate periodic clustering, the clustering algorithm should construct clusters with minimal communication overhead. Hence we recommend a passive clustering approach.

CHs often transmit data over long distances, hence they lose more energy compared to member nodes. It is essential to rotate the role of CHs among nodes so as not to burden a few nodes with more duties than others. The network is reclustered periodically in order to select energy-abundant nodes to serve as CHs, thus distributing the load uniformly on all the nodes. The passive clustering method may be used to reduce the overheads during clustering process, where the node that first proclaims becomes CH. In the proposed work, higher energy nodes are elected as CH to attain energy efficiency, hence the proclamation delay is defined as the function of node's residual energy. The selected CH can forward the aggregated data to the central base station through a multi hop backbone constructed only by the CHs. The proposed algorithm follows a deterministic approach for CH election, thus guarantees uniform distribution of CHs. The algorithm is distributed in nature, hence provides enough space for scalability.

3.2.1. Weight Calculation of Node

The election of the CH is done on the basis of the highest residual energy. This means that a node becomes a CH or a cluster member, depending on its own and one hop neighbor's residual energy. In WSN, there are several heuristics for selecting CHs. Residual energy, node degree, distance from the Base Station (BS), node ID and cumulative time for which the node acted as CH (node priority) are the prominent heuristics. Since the residual energy of the node is the energy aware heuristics, the proposed work uses it as the weighting parameter. Since CHs are overloaded with multiple tasks, the rate of energy depletion is also high. Hence the

CH elected should have high residual energy. Thus, for an energy efficient clustering approach, residual energy is the major indicator for the dominant set of nodes in both homogeneous and heterogeneous (energy) networks. Acquiring residual energy is again an internal task, doesn't need any communication. The weight of the node (W_e), is given as,

$$W_e(n) = E_{res}(n)/E_{init}(n) \quad (1)$$

Where, $E_{res}(n)$ is the residual energy of node 'n' and $E_{init}(n)$ is the initial energy of the node 'n'.

3.2.2. Passive Clustering

In passive clustering, Cluster head is selected based on "first declaration wins" rule, therefore the node that first proclaims becomes the CH. In the earlier works, the proclamation delay is random. Random delay may result in selection of low energy nodes as CH, hence decreases the energy efficiency of the system. In the proposed work, to elect most appropriate nodes, the proclamation delay is made inversely proportional to the node's weight. Once the proclamation delay expires, the node proclaims itself as CH. If a node hears a proclamation before the expiration of its proclamation delay, it refrains itself from the contention to become CH. The waiting time T_w of node n is given as

$$T_w(n) = k / W_e(n) \quad (2)$$

Where, k is a constant.
If a node receives multiple proclamations from different nodes, it selects the nearest node as its CH and associates with it, thus clusters are formed.

3.3. Exploiting Spatial correlation

In a densely deployed WSN nearer nodes sense similar data, due to their spatial proximity. When the spatially similar data are sent over the network, they consume substantial amounts of network's energy without any informational value. In a conventional clustered network, all the nodes in the cluster transmit their data to cluster heads, where the data is aggregated and sent to the sink. In a clustered aggregation scheme, clusters act as local filters on spatially redundant data, thus discard the further flow of insignificant data. However, suppressing the redundant data at the node itself, might be a better option than filters it out at the CH. Here in the proposed work, based on the time series sent by the sensors, the cluster head assigns them to different sub clusters based on their data similarity. For each sub cluster, an SCH node reports the data to the CH. Thus the spatially redundant data are filtered locally.
Sub cluster formation is done in three folds. At first, a high energy node is identified as SCH. Then, its closest neighbors are discovered. Finally, the data series of the neighbors are compared to the SCH's data series. The comparison is about the magnitude and trend similarity. If the neighbors are magnitude and trend similar with the SCH, the nodes are added to the sub cluster headed by SCH. From the rest of the cluster members, the next higher energy node is identified and the process repeats till all nodes inside the cluster are sub-clustered.

3.3.1. Selection of SCHs

Each sub cluster reports the data to the CH, through the SCH. Other nodes report only when anomalies occur. Thus the SCH selected should have higher residual energy than other nodes.

CH estimates the residual energy of all the cluster members and selects high energy node as SCH. Then the CH performs similarity estimation of nearby nodes with SCH. After identifying the sub cluster members for the SCH, CH adds them to the sub cluster headed by SCH. CH then identifies the high energy node from the rest. The process repeats till all nodes are assigned to a sub cluster.

3.3.2. Similarity Estimation

Once the SCH node in the cluster is identified, the geographical distance D is identified between the SCH and its neighbors. If the distance D is lesser than D_{th}, the nodes are said to be spatially similar. Here D_{th} is the maximum threshold distance for a sub cluster. As the nodes geographical locations are known to the cluster head, calculating the distance between the nodes is an easier task. Data similarity between the nearby nodes and SCH is assessed. The similarity identification is in two folds, first the magnitude similarity between the nodes is measured, and then the trend similarity is also estimated. The nodes that satisfy both the conditions are grouped into a sub cluster. The outlier nodes are considered for consecutive sub clustering rounds. From the rest of the nodes the next higher energy node is selected, its neighbors are listed and similarity is measured. This process continues until all the nodes in the cluster are sub clustered.

3.3.2.2. Magnitude Similarity

Let x (x_1, x_2,…, x_n) is the time series of representative node and y (y_1, y_2,…, y_n) is the time series of node y. Both the time series are of the same scale. The CH measures the Euclidean distance between the two time series that indicates the magnitude similarity between the time series. Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two time series. The Euclidean distance is a fair measure of similarity, since it compares the relationship between actual readings.

The Euclidean distance between the two time series is given by

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

The two time series are magnitude similar, if d (x, y) < α, where α is the user defined magnitude similarity threshold.

3.3.2.3. Trend Similarity

The Euclidean distance fails to identify the trend information of two time series; hence we measure the trend similarity between time series using cross correlation estimations. The correlation coefficient is a measure of the degree of linear association between two time series. Here we consider the Pearson's correlation coefficient as an effective indicator of trend similarity between two time series. The Pearson's correlation coefficient between data series x and y is given as

$$T_{(x,\,y)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (4)$$

Where $\bar{x}$ and $\bar{y}$ are the mean values of data series from SCH and neighbor node y. The value of $T_{(x,\,y)}$ ranges from -1 to 1. Where +1 indicates perfect positive correlation, -1 indicates perfect negative correlation between the two time series and zero indicates no correlation.

3.3.3. Spatial Data Suppression

The data series are said to be α similar, if d (x, y) < α and T (x, y) > 0.9. If the above conditions are met, the CH now adds the node y to the sub cluster represented by x. Sub clusters' are created, such a way that all its members are α similar with the sub cluster representative. Since all the members are α similar with the representative, only the representative data is communicated to the CH, the rest can be approximated within the data error bound by α. This scheme substantially reduces the data communications within the cluster. Consequently reduces the overall energy expense of the sensors. The algorithm for sub clustering is indicated below, where W is the total nodes inside a cluster, V_{nbr} is the neighbors of node V and V_{sc} is the Sub cluster headed by node V.

Algorithm.1. Formation of Sub clusters

1. Generate 1-hop cluster of W nodes
2. **While** (W>0)
3. V=highest energy node (W)
4. W-=V;
5. V_{nbr} Є D(V,W)<D_{th})
6. **For all** (V_{nbr})
7. **If** euclidean dist (V, V_{nbr}) < α
8. **If** correlation coeff(V,V_{nbr}) > 0.9
9. Add V_{nbr} in to V_{sc}
10. W-=V_{nbr};
11. **End if**
12. **End if**
13. **End for**
14. **End while**

3.4. Exploiting Temporal correlation

Temporal correlation is the measure of similarity between consecutive observations of node over the time. Most of the environmental variables show a good level of temporal correlation, due to their slow varying nature. This temporal correlation adds up a significant amount of redundant data over the time span. To reduce energy consumption, temporal correlation among data is exploited to identify a subset of sensor readings from which the remaining measurements can be predicted within the user defined accuracy. Readings which can be predicted from already delivered data need not be reported to the base station, thus communication is reduced. Digital filters are defined to interpret the sensor data in time domain, to predict the future values. The short term linearity of the signal is estimated using an LMS filter. Based on the estimation, future data are predicted as a linear combination of recent data history.

3.4.1. Prediction based Reporting

In a clustered data aggregation scheme, data reduction is achieved by employing prediction based reporting, where the prediction processes take place at both the sensor node and the CH

simultaneously using identical filters. In the sensor node, at each sampling instant t, the actual sensed value is compared to the value predicted by the model. If the difference between the two is lesser than the threshold, no data is transmitted. If the difference is higher than the threshold, the data are transmitted to the CH. As a result only a fraction of data is transmitted.

In the DPF, there are three distinct modes of operation: initialization, normal and stand-alone. A node goes through the initialization mode only at the beginning and then switches between the normal and standalone modes. During the initialization mode, at every sampling instant t, the sensor node transmits the observed data to the CH. A prediction model is constructed by the sensor node in parallel. The energy consumption of the mode is given as

$$E_{init} = E_{tx} + E_{rx} + E_{pred} \tag{5}$$

Where E_{tx} and E_{rx} are the transmission and reception energy of sensor and CH node respectively. E_{pred} is the prediction energy incurred by the sensor node during model construction.

The prediction is said to be converged, if the prediction error is lesser than the error threshold β for M consecutive predictions. Now the sensor node switches to standalone mode, by communicating the prediction model to the CH. In standalone mode, at each sampling instant t, the sensor node and CH predict the data using the prediction model based on data history. Along with prediction, sensor node still collects data and compares the actual sensed value with the value predicted by the filter. If the deviation is lesser than the threshold β, the filter model is assumed to be accurate for that time instant. Hence the filter is fed with the prediction y [t] instead of sensed value x [t]. Similarly, the sink predicts a value based on model and uses it as an approximation of the actual observation for the time instant. During the stand alone mode, the energy is saved as the sensor node does not report its readings to the CH. The energy consumption of the dual prediction is given as

$$E_{sa} = 2E_{pred} \tag{6}$$

If the error exceeds α, the mode switches to normal mode. During normal mode, the data are transmitted to the CH. The prediction engine at the sensor node adjusts the weight values towards the convergence of the prediction with the desired value. Once the prediction is converged, the mode switches again to standalone mode. The energy consumption during normal mode is given as

$$E_{norm} = E_{tx} + E_{rx} + E_{pred} \tag{7}$$

3.4.2. LMS based Prediction filter

In the proposed work the prediction filter is built on LMS algorithm. The functional features of an LMS based prediction filter are briefed in this part. A linear adaptive filter samples a data stream x over a length n at an instant k, which is denoted as x [k] and calculates a prediction as $y[k] = w^{T}[k] \cdot x[k]$, which effectively is a linear combination of the previous n samples of the data stream, weighed by the corresponding weight vector w [k]. The output y [k] is then compared to the desired signal d [k]. The functionality of the LMS based prediction filter is shown in **Fig. 2**.

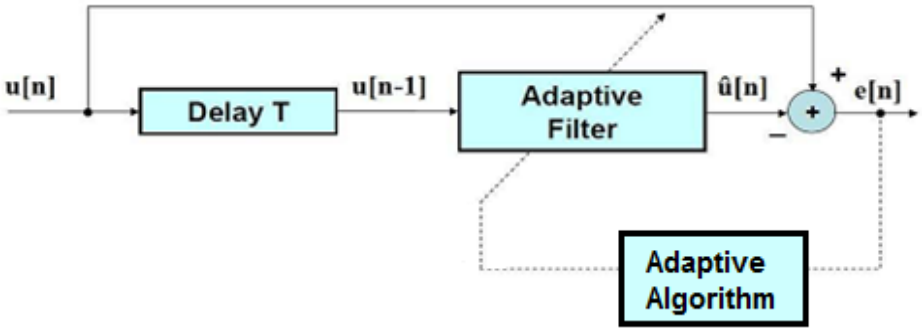

Fig. 2. LMS based Prediction filter

The prediction error e [k] is then computed as: e [k] = y [k] − d [k] and fed into the adaptation algorithm, so the filter weights are updated at each time step k in order to minimize the mean square error. In the normalized LMS, the step size is normalized on every time step, thus the sensitivity to the input signal is reduced. The functional model of nLMS algorithm is defined in the **Table 1**.

Table 1. LMS model

$y[k]=\underline{w}^T[k]\underline{x}[k]$	Filter output
$e[k]=d[k]-y[k]$	Estimation Error
$\underline{w}[k+1]=\underline{w}[k]+\mu\underline{x}[k]e[k]/\underline{x}^T[k]\underline{x}[k]$	Weights adaptation

3.4.3 VSS-nLMS prediction

Step size plays a crucial role in achieving data accuracy and energy efficiency in the prediction based reporting. There is no specific optimum value for step size, since it is context dependent. When the deviation is high, larger step size attains faster convergence. Near the point of convergence smaller step sizes realize steady state prediction. The work adapts the step size with the state of prediction and achieves a considerable speed of convergence and reduced deviations. Here we introduce an integer D that controls the step size during different states of prediction.

The change in weight is made in multiples of the step size μ, whose value ranges between 0 and $1/E_x$, where E_x is the mean input power.

$$E_x = \frac{1}{M}\sum_{k=1}^{M} x[k]^2 \tag{8}$$

$$\mu = (1/E_x)/D$$

D = $D_{max,}$ during steady state,
D = $D_{min,}$ during convergence state.

3.5. Exploiting spatio temporal correlation (Collective Prediction)

In a conventional prediction based reporting approach all the cluster members report their data to CH through independent prediction filters. In the proposed work, a sub cluster is attributed by highly correlated data sources located in close vicinity and are represented by the single SCH node. Thus the SCH node's prediction filter is considered sufficient for the whole sub cluster. Here the prediction based reporting is between the SCH and CH using a common model. Since the representative node has a close correlation with its neighbors, the data generated by representative node resembles the entire sub cluster in both magnitude and trend. Thus, by collecting only the representative data, entire sub cluster's data can be approximated within an error bound of α. Each representative node constructs a prediction model based on the data observed by it on behalf of the sub cluster. The model and its parameters are transmitted to CH and other sub cluster members.

The sub cluster member executes the model as follows: at each sampling instant t, the actual sensed value is compared to the value predicted by the model. If the difference between the two is lesser than the threshold, then the model is assumed to be β similar to the node data for that time instant. Let $\beta=k\delta_i$, where k is an application specified real constant larger than 1. The function δ_i is the standard deviation of the white noise in the prediction system, and it also provides a measurement of the accuracy of the prediction. Let x_i (t) be the actual sensor reading of sensor node i and p_i (t) be the estimated value of x_i (t) at time t stored at the cluster head, then x_i (t) Є $[p_i (t) - \beta, p_i (t) + \beta]$ with an error probability at most $1/k^2$.

The CH can compute the value through prediction model and uses it as α approximation of the actual observation of sub cluster member nodes. Thus, so long as the model predictions are similar to the observed values, no communication is necessary between the sub cluster members and the CH. In contrast, if the difference between the sensed data and the model predicted data exceeds a threshold, the sensed value is pushed to the CH.

The similarity between the sensor readings of the two sensor nodes can be expressed using their Euclidean distance. The Euclidean distance between the readings of any two sensor nodes S_i and S_j in S at time t is defined as $d_t(i,j) = |x_i(t) - x_j(t)|$. From eqn. 4 the trend similarity between the time series is estimated. When the data series are closely correlated, the Euclidean distance between the individual elements of the data series remains unchanged. Here we estimate the real sensor readings of sensor node S_i by utilizing the local estimation of sensor node S_j. This estimation of x_i (t) through p_j (t) introduces errors in both spatial and temporal domains. The spatial error is due to the deviation of the Euclidean distance between x_i (t) and x_j (t). The temporal error is the local estimation error between $x_j(t)$ and $p_j(t)$. Let E_{ij} (t) be the estimation error of estimating x_i (t) by p_j (t).

$$\begin{aligned} E_{ij}(t) &= |x_i(t) - p_j(t)| \\ &= |x_i(t) - x_j(t) + x_j(t) - p_j(t)| <= |x_j(t)-p_j(t)| + |x_i(t)-x_j(t)| \\ &= e_j(t) + d_t(i,j) \end{aligned} \quad (9)$$

$$Max(E_{ij}(t)) = max(e_j(t)) + d_t(i,j) \quad (10)$$

$$Max(E_{ij}(t)) = \beta + d_t(i,j) \quad (11)$$

Thus, we have E_{ij} (t) Є $[0, \beta+d_t (i, j)]$. As x_i (t) Є $[p_i (t) -\beta, p_i (t) +\beta]$ with error probability at most$1/k^2$, it is easy to see that the estimation error E_{ij} (t) Є $[0, \beta+d_t (I, j)]$ with error probability at most$1/k^2$. Sensor node S_i is Δ- similar to S_j at time t, if and only if $|e_i(t)+d_t(i,j)| <= \Delta$, where Δ is a positive real constant.

Thus, the sub cluster members push data only when the observed data deviates from the value predicted by the common model, thereby capturing deviating trends. The state diagram of the CoPeST framework is shown in **Fig. 3**. The proposed system employs VSS-nLMS based prediction filter for constructing prediction models, which is computationally inexpensive and provides optimum level of accuracy. Our approach incorporates active feedback between the CH, representative and other sub cluster members' results in high reliability of data with considerable energy conservation.

The CH initiates sub cluster adjustment by measuring the spatial deviations over the sliding window. On each window, the total collected data comprises of predicted data and SCM updated deviations. Inside the window, if the deviation is greater than predicted data, the sub cluster is considered invalid. The cluster head puts the deviating node in to another closer sub cluster or assign it to a separate sub cluster.

4. Experimental Classification Results and Analysis

The very purpose of implementing CoPeST is to achieve an energy efficient data collection. The energy efficiency is measured in terms of reduced number of communication packets. The approach achieves energy conservation at the cost of marginal tolerance in the data accuracy, hence it is essential to analyze the average error of data acquired at the base station with respect to the original data observed at the sensors. There are numerous measures for estimating the data error in the distributed data gathering system. Here the data error is measured in terms of mean absolute error (MAE) of the received data. The performance of the proposed work CoPeST is evaluated on MATLAB platform. In order to investigate the performance of CoPeST with large-scale networks, we generate large traces of a spatially correlated data set based on a mathematical model proposed in [21], through which the model parameters are extracted from small-scale real data sets [22]. The work is evaluated by comparing the performance with other energy efficient data gathering approaches. Then the impact of spatial and temporal error threshold on the performance is evaluated. The cluster size and its impact on the performance is also analyzed. The scalability of the system is also analyzed.

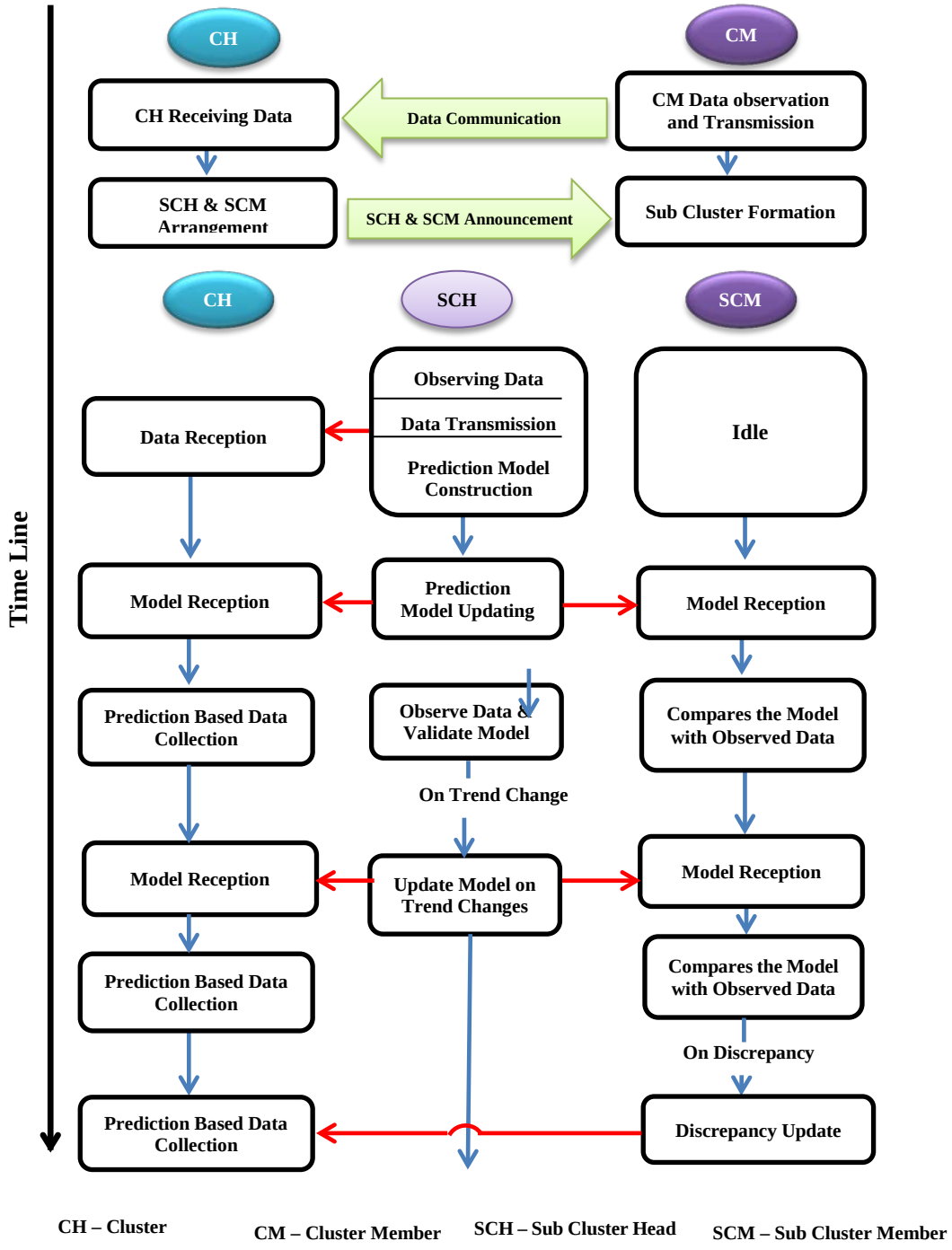

Fig. 3. Copest flow diagram

4.1 Comparison with other approaches

The work is evaluated by the amount of data reduction, for different error thresholds (Δ). With the increased Δ, the communication cost is reduced. The proposed work CoPeST is compared with other data reduction approaches like PRESTO and ASAP. PRESTO approach involves model driven push, where the temporal correlation of the sensor data is used for data reduction. In ASAP, the spatial correlation between the data is used to select a fraction of nodes to report the data to the CH and the rest are predicted. The former exploits only temporal correlation and the latter uses only spatial correlation. In ASAP, the error threshold is considered as a spatial error threshold, in PRESTO, the error threshold is considered as a temporal error threshold. In CoPeST, the error threshold (Δ) is split into two equal error thresholds namely temporal error threshold (α) and spatial error threshold (β). (Δ = α + β). The data reduction achieved by CoPeST outperforms both PRESTO and ASAP, since it jointly exploits spatio-temporal correlation among the sensor data.

In CoPeST, the data reduction is in two folds. First the number of reporting nodes is reduced. This small portion of nodes also sends only a fraction of observed data. For small Δ values, ASAP and CoPeST achieves lower message cost than PRESTO. Since CoPeST bifurcates the error threshold, during low Δ values, spatial and temporal models experience tight constraints. This necessitates frequent updates to ensure the data within the specified error threshold. At Δ=0. 2°C, CoPeST can reduce only 40% data. When Δ increases, the temporal data reduction is steep along with moderate reduction in active reporting nodes. Thus the dual reduction approach performs well and CoPeST outperforms both PRESTO and ASAP. At Δ=1°C, CoPeST can reduce about 75% data. The message comparison is shown in Fig.4.

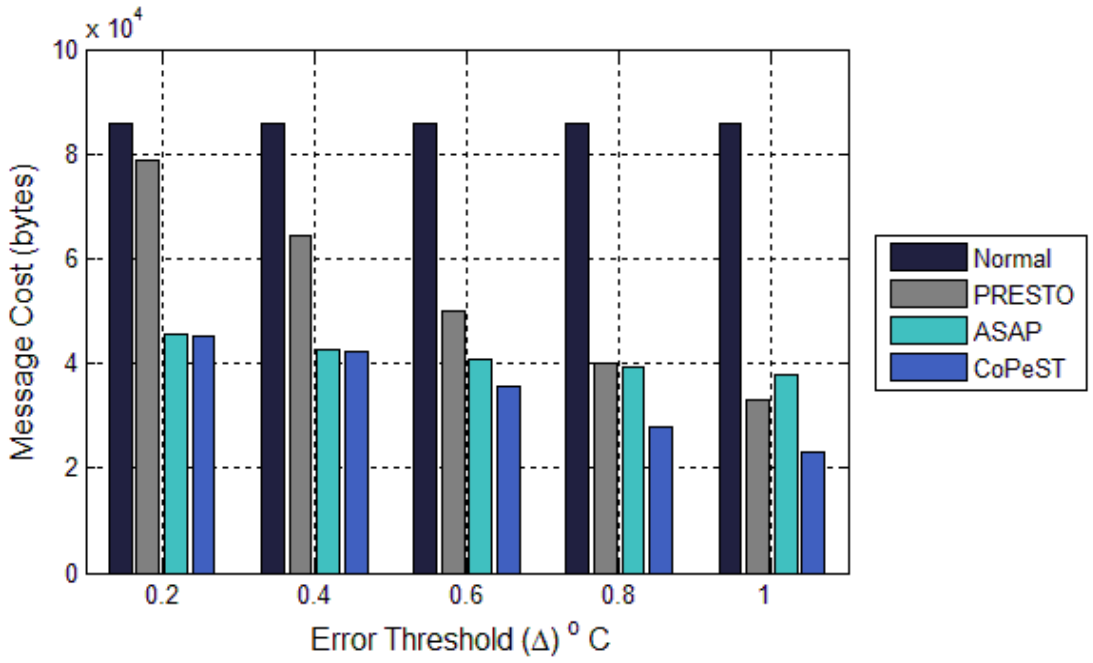

Fig. 4. Comparison of message costs (CoPeST, ASAP and PRESTO)

The energy model of TELOSB [4] specified in section I is used for energy estimation of our work. In CoPeST, (5N+5) cycles are required for each prediction. For the filter length of 4,

the energy cost of prediction per round is 30nJ.Each round, the energy cost of transmitting and receiving a 16 bit data is 11.52μJ and 12.59μJ respectively. From eq. (5-7), the energy cost during initialization and normal modes is 24.14μJ and during standalone mode is 60nJ.

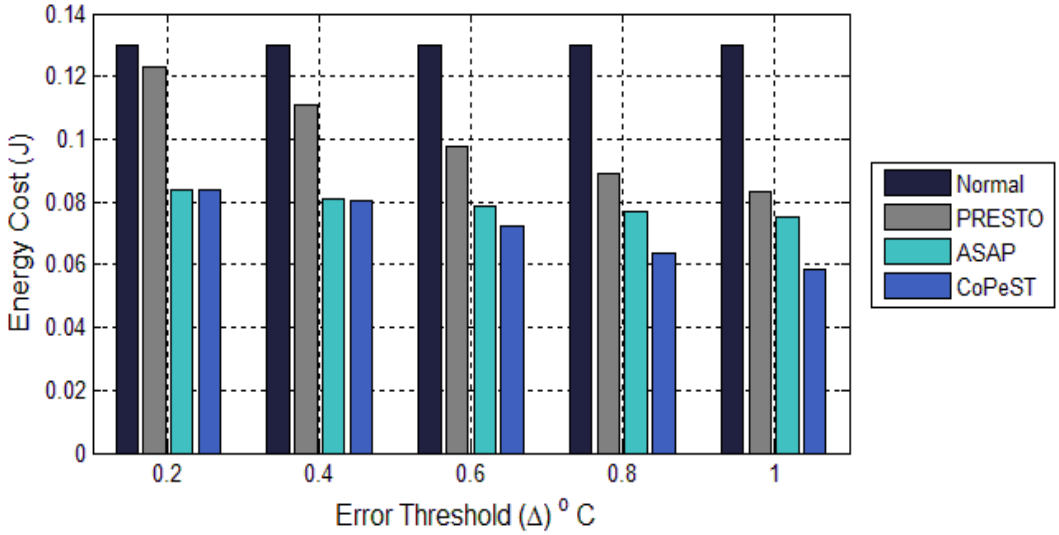

Fig. 5. Comparison of energy costs (CoPeST, ASAP and PRESTO)

The **Fig. 5** shows the comparison of the energy cost of different protocols for various error thresholds. The energy consumption of CoPeST is 50% lesser than that of the conventional data gathering at higher error thresholds.

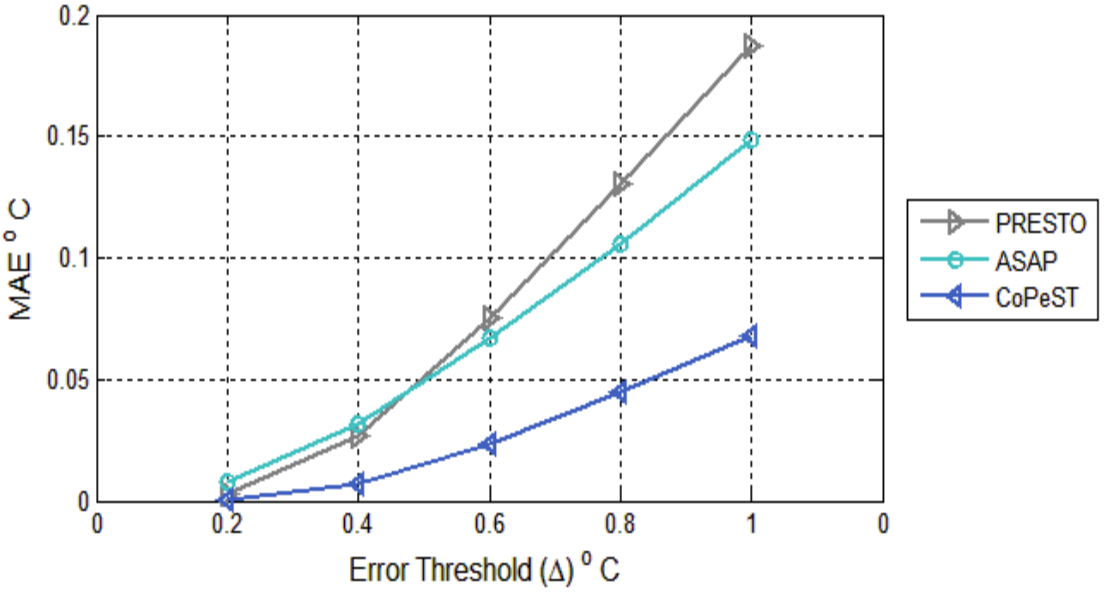

Fig. 6. Comparison of mean absolute errors (CoPeST, ASAP and PRESTO)

Another metric to evaluate the performance of a data gathering approach is to analyze the mean absolute deviation between the observed data and the data collected at the base station. The effectiveness of system towards reducing the deviations is measured using this performance. In the proposed system, data deviation is due to two main factors. One is the deviation in the prediction model, due to trend changes in the sensor measurement. Secondly the deviation between the representative's data and the sub cluster member's observation due to spatial distortions. The deviation trend with the increased Δ is smoother for CoPeST and increase abruptly for the ASAP and the PRESTO. Compared to PRESTO and ASAP, the mean deviation is much lesser in CoPeST, due to the bifurcation of error threshold into spatial and temporal error thresholds and overlapping of spatial and temporal errors. As in **Fig. 6** for higher Δ values, the PRESTO has higher mean deviations, since high temporal error threshold allows the prediction to deviate for larger values. At Δ=1°C, the deviation of CoPeST is less than a half of the ASAP and PRESTO.

4.2 Impact Of Spatio-Temporal Error Threshold

Since the system involves both spatial and temporal data reduction, we analyze the impact of α and β discretely on the performance of the proposed system. The temporal error threshold α decides the frequency of trend change updates. Higher the , lesser the frequency of updates and vice versa. The β decide the amount of active nodes to report the trend changes in the network. The low value of β increases the number of active nodes, which increases the spatial granularity of the data observation. The higher value of β reduces the active reporting nodes, thus conserves substantial amount of energy. The right combination of α and β better trades off between the data accuracy and energy conservation. The α value is incremented from 0.2°C to 0.8°C and for each α value, β is varied from 0.2°C to 0.8°C. For every combination of α and β, the message costs and mean data deviations are estimated. From the **Fig. 7**, it is observed that the impact of spatial error threshold of data reduction is smooth, but the impact of temporal error threshold of data reduction is sharp. When we increase the spatial error tolerance the reduction of active nodes is limited by distance threshold and the maximum number of nodes in the cluster. Thus, increased spatial error tolerance cannot further reduce the number of active nodes.

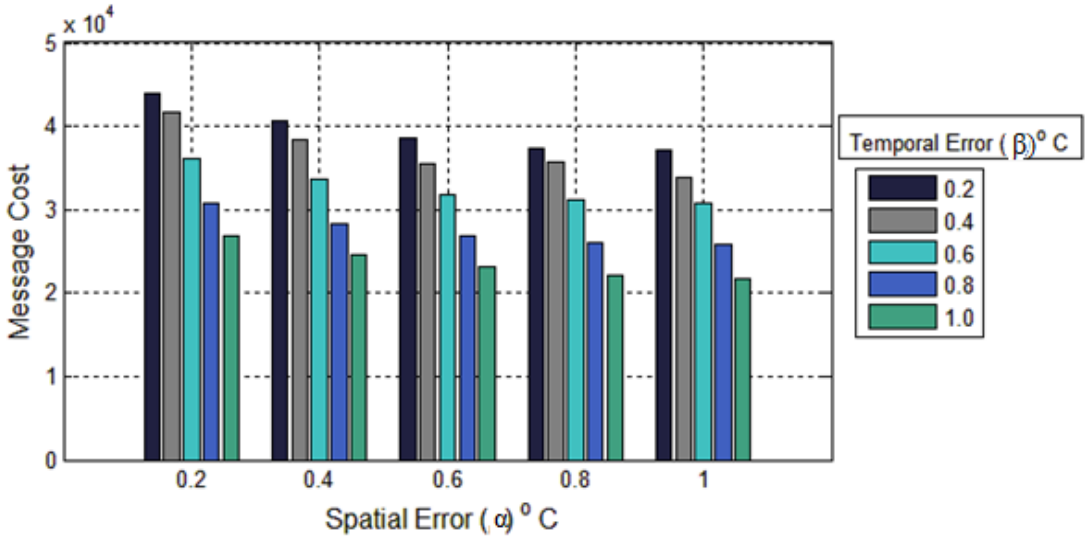

Fig. 7. Message costs for various spatial and temporal error thresholds

The system combines the advantages of both spatial correlation and temporal correlation among sensor data, hence it is essential to evaluate individual performance on message cost to evaluate their importance. Here we show the separate message costs of temporal correlation based data collection (dual prediction based reporting) and spatial data deviance updates (collective data push) for different combinations of α and β. The temporal correlation based data reporting is between the representative node and the CH. The spatial data adjustments are among the sub cluster members and the CH. This evaluation helps in identifying the right combination of α and β to achieve an efficient data collection. From **Fig. 8**, it is observed that larger temporal and smaller spatial error threshold is the wise choice to achieve significant data reduction along with optimal accuracy on collected data.

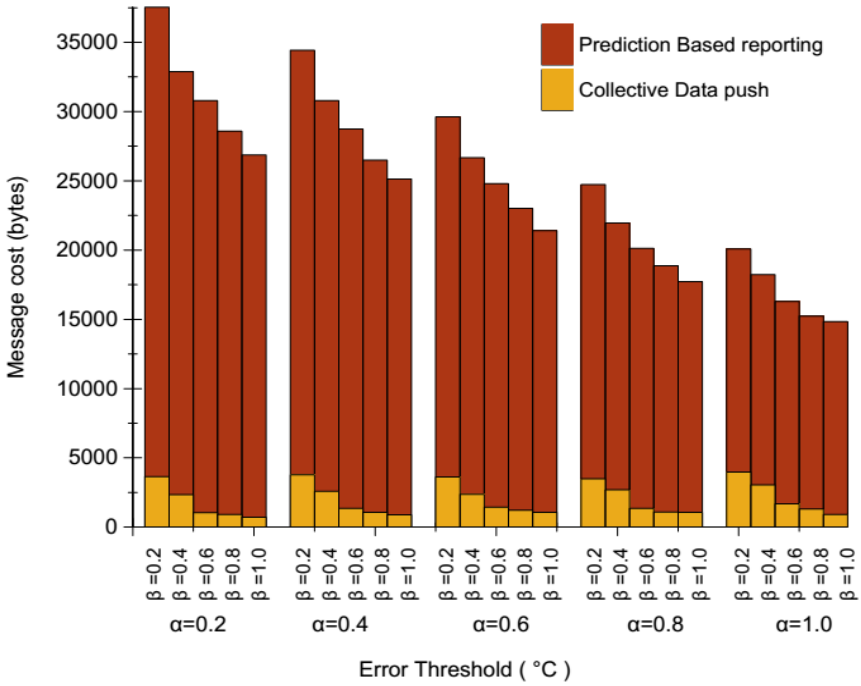

Fig. 8. Discrete message costs for various spatial and temporal error thresholds

In the same way, the effects of α and β are analyzed on the mean absolute error of the collected data. With the increase in α, the MAE also increases. The spatial observation error increase with the increased β. The data deviation is inversely proportional to the data reduction. As in the case of data reduction, data deviation increases sharply with increased α and increases gradually with increased β as shown in **Fig. 9**.

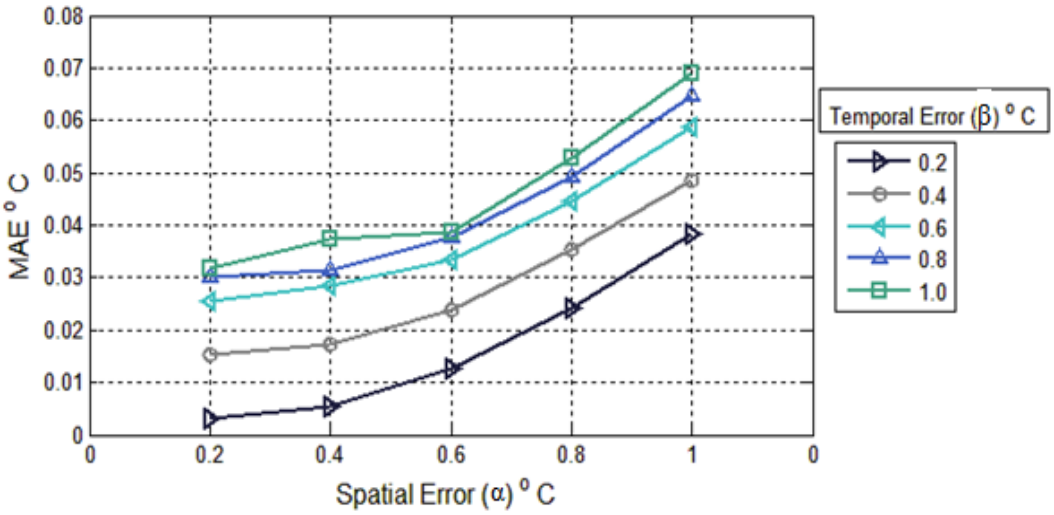

Fig. 9. MAE for various spatial and temporal error thresholds

4.3 Impact of Cluster size

Here we analyze the impact of cluster size on the efficiency of our proposed system. The cluster size decides the number of nodes in a cluster. The larger the cluster size brings in more nodes, hence the nodes per sub cluster also increases. The heavy weight sub clusters reduces the number of active reporting nodes, consequently improves the energy efficiency. The data reduction and accuracy of the framework with different cluster sizes are indicated in **Fig. 10**.

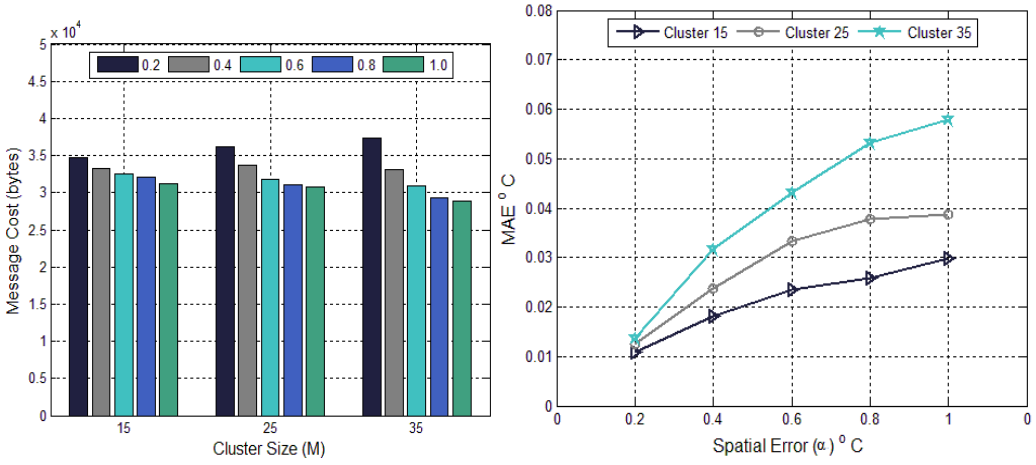

Fig. 10. Impact of cluster size on (a) message cost (b) Mean absolute error

The increased cluster size helps in achieving more spatial reduction. In a small cluster size, the spatial data reduction with respect to increased spatial error threshold is smooth. The data reduction for 0.2°C tolerance is 42K and 0.8°C tolerance is 38K. But in a larger cluster, there is a steep reduction in data communication with respect to spatial error tolerance. The data reduction for 0.2°C tolerance is 46K and 0.8°C tolerance is 29K. The increased cluster size

brings a double benefit for the proposed data reduction approach. Since the number active nodes are less, the total trend updates are also reduced. This further decreases the nodal communications. From the simulations, it is observed that the cluster size has no direct impact on the temporal data reduction, but has high impact on the spatial data reduction. The mean data deviation is also analyzed for different cluster sizes. Increased spatial error threshold improves the data reduction at the cost of increased data deviation. In small clusters, the difference in data deviation with respect to increased spatial error threshold is minimal. In large clusters, when the spatial threshold increases, more nodes are put into passive mode. Therefore the data deviation also increases abruptly.

4.4 Scalability

In a distributed data gathering approach, the scalability is an important parameter. The proposed work is evaluated on networks of various scales. The performance has improved with the size of the network. Increased number of nodes increases the node density of the network. When node density increases more nodes get into close proximity, results in a significant increase in spatially correlated data. This close proximity increases the size of sub clusters. **Fig. 11**. shows the increased size of the network exponentially increases the number of sub clusters in the network, hence the percentage of active nodes is reduced.

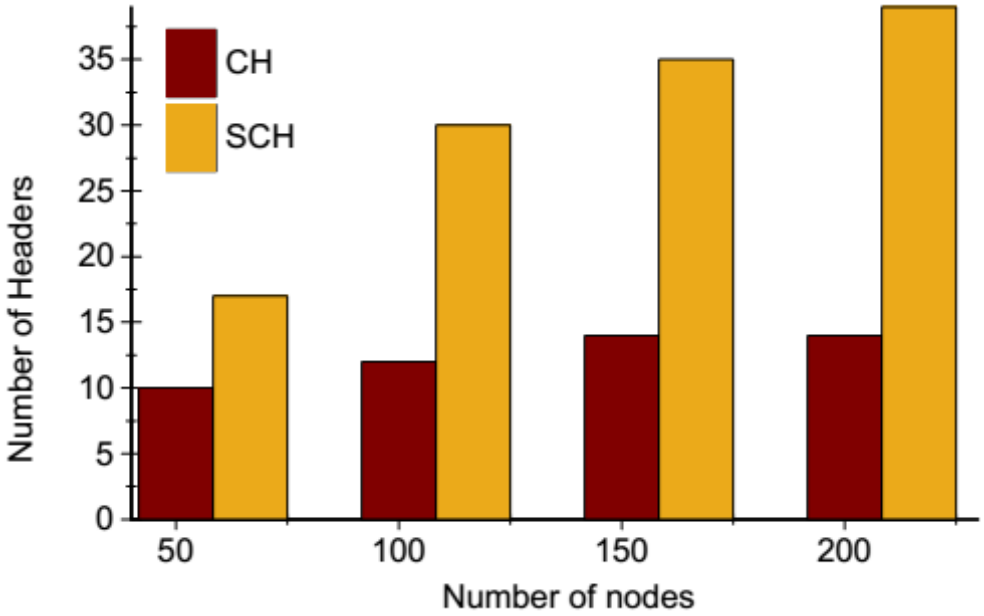

Fig. 11. Number of Cluster and Sub cluster heads for different network scales

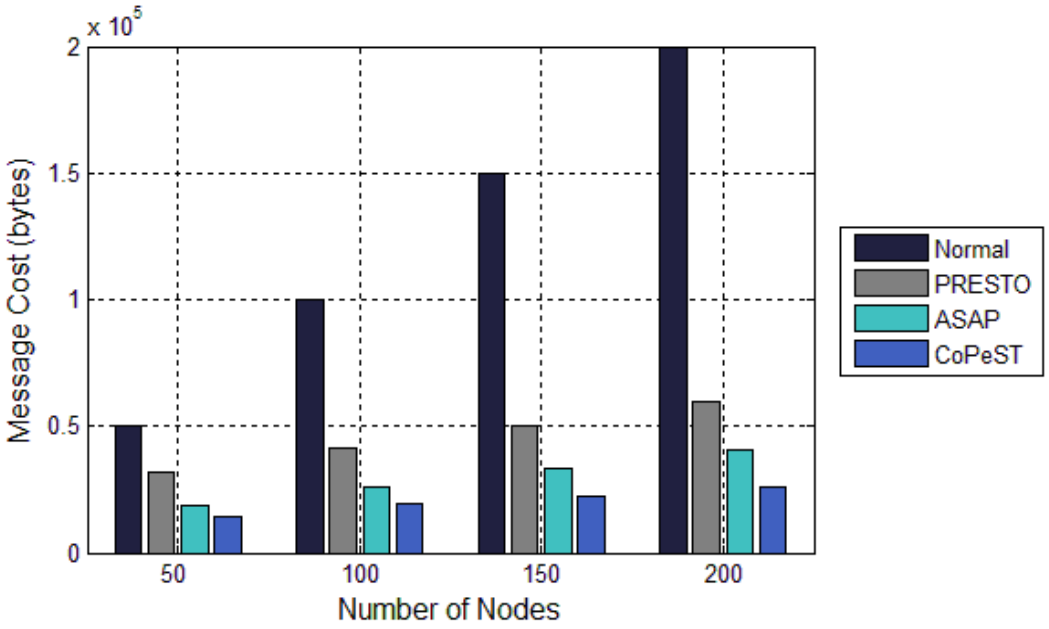

Fig. 12. Message cost comparison for different network scales

The **Fig. 12**. shows the proposed work suits well with the large scale networks. With the increase in number of nodes, the data load of the network increases. Here in the proposed work, the spatially redundant data are filtered out by the sub cluster based reporting. This is the reason for the major data reduction in the system. From **Fig. 12**, it is inferred that with the increase in node density, CoPeST increases the percentage data reduction from 75% to 87%.

5. Conclusion

CoPeST achieves two level data reduction without much deviance in the collected data accuracy. The work has been evaluated with the amount of data reduction and mean absolute data deviation. Since the work jointly reduces the spatially and temporally redundant data, the data reduction has improved to multiple folds than the previous works. It is highly a supervised mechanism that guarantees user specified error threshold in spatial and temporal aspects. The system's performance improves with the increased error tolerance. The system has proven to be highly scalable. The impact of various parameters is analyzed in detail. The system has reduced the data transmission up to 75% with a mean absolute deviation of 0.07°C. The future work involves dynamically adjusting the spatial and temporal error thresholds based on data dynamics and spatial variations.

References

[1] M. Li, Y. Liu and L. Chen, "Non threshold-based event detection for 3D environment monitoring in sensor networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no.12, pp. 1699-1711, December, 2008. Article (CrossRef Link)

[2] G. Li, J. He and Y. Fu, "Group-based intrusion detection system in wireless sensor networks," *Computer Communications*, vol. 31, no. 18, pp. 4324-4332, December, 2008. Article (CrossRef Link)

[3] Mo Li and Yunhao Liu, "Underground coal mine monitoring with wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 5.2, no. 10, March, 2009. Article (CrossRef Link)

[4] G.D.Meulenaer, F.Gosset, O.Standaert and O.Pereira, "On the energy cost of communication and cryptography in wireless sensor networks," in *Proc. of IEEE International Conference on Wireless and Mobile Computing*, October 12-14, 2008. Article (CrossRef Link)

[5] M. Srivastava, David Culler and Deborah Estrin, "Guest editors' introduction: Overview of sensor networks," *Computer*, vol. 37, no.8, pp. 0041-49, August, 2004. Article (CrossRef Link)

[6] D. Chu, A.Deshpande, J.M.Hellerstein and W. Hong, "Approximate data collection in sensor networks using probabilistic models," in *Proc. of the 22nd International Conference on Data Engineering*, April , 2006. Article (CrossRef Link)

[7] G.Anastasi, M.Conti, M.D.Francesco and A.Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad hoc networks* 7.3, pp. 537-568, May, 2009. Article (CrossRef Link)

[8] Supriyo Chatterjea and Paul Havinga, "An adaptive and autonomous sensor sampling frequency control scheme for energy-efficient data acquisition in wireless sensor networks," *Distributed Computing in Sensor Systems*, pp. 60-78, June 11-14, 2008. Article (CrossRef Link)

[9] C. Carvalho, D. G. Gomes, N. Agoulmine and J. N. de Souza, "Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation," *Sensors*, vol. 11, no. 11, pp. 10010-10037, October, 2011. Article (CrossRef Link)

[10] Guorui Li and Ying Wang "Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*,vol. 1, pp. 1-13, March, 2013. Article (CrossRef Link)

[11] Iosif Lazaridis and Sharad Mehrotra, "Capturing sensor-generated time series with quality guarantees," in *Proc. of 19th International Conference on Data Engineering, IEEE*, March, 2003. Article (CrossRef Link)

[12] Ming Li, Deepak Ganesan and Prashant Shenoy, "PRESTO: feedback-driven data management in sensor networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 4 ,pp. 1256-1269, August, 2009. Article (CrossRef Link)

[13] Silvia Santini and Kay Römer, "An adaptive strategy for quality-based data reduction in wireless sensor networks," in *Proc. of the 3rd international conference on networked sensing systems*, May, 2006. Article (CrossRef Link)

[14] L.Villas, A.Boukerche, H.Oliveira, R.Araujo and A.Loureiro, "A spatial correlation aware algorithm to perform efficient data collection in wireless sensor networks," *Ad Hoc Networks*, vol. 12, pp. 69-85, January, 2014. Article (CrossRef Link)

[15] F.Emekci, S.E.Tuna, D.Agrawal and A.E.Abbadi, "Binocular: a system monitoring framework," in *Proc. of the 1st international workshop on Data management for sensor networks: in conjunction with VLDB*. ACM, August 30, 2004. Article (CrossRef Link)

[16] Bugra Gedik, Ling Liu and Philip S.Yu, "ASAP: an adaptive sampling approach to data collection in sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 12, pp. 1766-1783, December, 2007. Article (CrossRef Link)

[17] Ian F. Akyildiz, Mehmet C.Vuran and Özgür B. Akan, "On exploiting spatial and temporal correlation in wireless sensor networks," in *Proc. of WiOpt*, vol. 4, March 24-26, 2004. Article (CrossRef Link)

[18] A.Deshpande, C.Guestrin, S.Madden, J.Hellerstein, and W.Hong, "Model-driven data acquisition in sensor networks," in *Proc. of the Thirtieth international conference on Very large data bases*, vol. 30, August 29-September 3,2004. Article (CrossRef Link)

[19] Chong Liu, Kui Wu and Jian Pei, "An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no.7, pp. 1010-1023, July, 2007.Article (CrossRef Link)

[20] Antonios Deligiannakis and Yannis Kotidis, "Exploiting spatio-temporal correlations for data processing in sensor networks," *GeoSensor Networks*, Springer Berlin Heidelberg, pp. 45-65, 2008. Article (CrossRef Link)
[21] Apoorva Jindal and Konstantinos Psounis, "Modeling spatially-correlated sensor network data," *First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, IEEE*, October 4-7, 2004. Article (CrossRef Link)
[22] [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

Muruganantham ARUNRAJA completed his Bachelor degree in Shanmugha college of Engineering, Thanjavur and completed his Master's degree in Embedded system technologies in Anna University, Thrunelveli. He has eight years of Industrial experience in Embedded systems and related areas. He is currently a full time research scholar in Anna University Regional office, Madurai. His areas of interest are embedded systems, wireless networks and instrumentation.

Veluchamy MALATHI is working as professor in the department of Electrical and Electronics Engineering in Anna University Regional office, Madurai. She completed her Bachelor degree in College of Engineering Guindy and her Masters in Thiyagaraja College of Engg, Madurai. She completed her Ph.D in Anna University Chennai and her areas of interest are intelligent techniques and its applications, Smart Grid, FPGA based power system and Automation.