IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AN ADVANCED SPATIO TEMPORAL ADAPTATION MODEL (ASTAM) FOR E-COMMERCE ENVIRONMENTS

Abraham Amal Raj B[1], Mahaveer Sain[2]

([1]Department of Computer Science & Informatics, Maharishi Arvind University, Jaipur, Rajasthan)
([2]Associate Professor, Department of Computer Science & Informatics, MAISM, Jaipur, Rajasthan)

ABSTRACT:

Wireless Sensor Network's (WSN) enactment and consistency essentially hinge on data redundancy. Sensory data's inherit assets are Spatial and temporal similarity. A considerable amount of nodal energy and bandwidth can be preserved by reducing this spatio-temporal data redundancy. To diminish data redundancy virtually all the data gathering methodologies use either spatial correlation or temporal correlation. The mistreatment of both the temporal and spatial correlation between sensory data is done in Spatial, temporal correlation with Advance Multiple data prediction Interface (SAMI), by similarity based sub clustering the spatial redundancy of sensor data is abridged in the projected work. A solitary illustrative node can represent a meticulously correlated node by similarity based sub clustering. By means of a model based prediction methodology temporal redundancy can be abridged, which helps in diffusion of only a subset of sensor data while the cessation is predicted. Reduction of a considerable amount of energy expensive communication is carried out in this scheme, while the error threshold of the data is user defined. The projected work is highly ascendable, as it is a distributed methodology. This work helps in achieving up to 75% reduction in data in an occasionally gathered system data where an error tolerance of 0.6°C is maintained.

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

I. INTRODUCTION

With the encroachments in MEMS, chip incorporation and Radio frequency expertise, Assortment of applications use WSN comprising environmental monitoring [1], military surveillance [2], industrial process controls [3], smart spaces and many more. For upholding required accuracy, we go for distribution of extensive term data gathering in WSN. Each sensor node in a WSN behaves as an identity enclosed system that encourages sensing, communicating and computing elements. Finite energy source is the major constraint of the sensor nodes. One of the primary energy consuming functionality is wireless communication, depending on the particular type of sensing performed it becomes an important role in various aspects. By other means, computation has been considered as least energy consuming activity. To achieve a longer lifespan by maintaining minimum energy consumption is an important objective of the excessive collection of data in deploying WSN, For enabling meaningful analysis high resolution and high quality are maintained sufficiently. The cost of the distributed monitoring depends mainly upon the number of active sensor nodes and quantity of sensor nodes transmitted data.

In a collection of periodic data methodology [4], to obtain a finest data granularity the nodes first senses the surrounding environment and then the data gets incessantly transmitted out of interest over a particular period of time. This observed data helps in enabling the data analysis which is highly composite. This process may be refined with high expenditure when the data are collected continuously and the WSN life span also gets reduced. In a constrained bandwidth, introduction of a number of nodes leads to disproportionate communication and this prevails heavy traffic where an excessive amount of data gets collided resulting in the reduction of the throughput. Redundant data constitutes a huge proportion of the total amount of transmitted data [5]. Without having any informational value, the redundant consumes a sustainable amount of the network resources. Auspiciously, there may be an aggressive reduction in data transmission for saving energy with no loss in large degradation of dependability observation; this may lead to the Spatio-temporal correlation existence in the sampling data. If the sensor nodes are close to each other, there may be similarities in their observations, hence the value of the neighboring nodes predicted easily. Physically nearby sensors have high spatial correlation. To identify the similarity nodes with highest energy for the formation of cluster head *Turan's* theorem based on Extremal Graph Theory have been applied. By looking the similarities in the magnitude and trend of the generated data, we can easily predict that they are neighboring

nodes because their spatial correlation will also be high. In this correlated group, most of the data can be approximated and the reported data belong to only a subset of sensors. In a sensor node, if the successive data are more similar than their sampling frequencies are higher, resulting in a simple prediction of the future results from the previous data collected in the same node. Approximating the trend of the signal from the temporal correlation helps in the prediction of future data.

In the projected work, the inter sensor and intra sensor of both temporal correlation and spatial correlation gets exploited helps in reduction of communication expense with no loss in accuracy. Sensors in a cluster having similar observations are grouped in various sub clusters. To represent the overall sub clusters any one of the sample nodes is selected from a sub cluster. Based on the collection of the previous data, the sample node accumulates a model of temporal correlation using LMS filter. The constructed model communicates with the other sub cluster members and the CH. By updating the appropriate filter coefficients the sample node updates the changes in trend. When the user defined threshold exceeds the observed data and the predicted data, the sub cluster members start transmitting the data. The periodic reporting framework's communication cost is effectively reduced by this system guaranteeing the accuracy of the user-defined nodes.

With numerous correlation degrees the projected system worth has been done on the synthetic data set. Better efficiency in this system is obtained in terms of energy. Best accuracy is obtained in the collected data. With no negotiation on data accuracy the nodal energy in the sub clusters has been balanced in an enhanced manner. In temporal prediction approach to identify the sub cluster head the idea of Large Subset given by *Turan's* have been implemented.

The rest of the projected work has been represented as follows: Section 2 brief discussion on the related works. Section 3 embellishment on the spatial and temporal correlation based data inference. We appraise SAMI by trace-driven simulations in Section 4. Section and 5 completes the paper.

II. RELATED WORK

In the strategy of wireless sensor networks energy efficient functionality is a key issue.For most of the WSN applications limited energy prevails as a bottleneck, based on energy maintenance numerous works have been going on energy maintenance in WSN. The key directions of energy maintenance in WSN have been discussed in Anatasi et al [6] where, duty cycling depends upon mobility and data driven and their methodologies are described. Redundant data can be reduced by data driven methodologies. Nodal energy maintenance can be obtained by reduction in the amount of data traffic, maintenance of bandwidth and avoidance of data collision.

The exploitation of the temporal correlation between successive data can result in the reduction of the temporal redundant data through multiple methodologies [7]. The communication overhead can be reduced by prediction of future date with the help of recent data history. Using linear regression methods [8] several data prediction methodologies, can exploit temporal correlations between sensory data, due to lack of adaptability in the dynamic variations of the input signal the accuracy also gets reduced. In [9], ARIMA based methods are used to predict the future sensor data from previous data history. One of the basic need of ARIMA is need of abundant basic data, which is computationally expensive and the prediction of the series becomes deprived when there are numerous turning points. In [10], the prediction had been executed using PCA that explains in detail about the prior model definition. In our projected work, we use LMS Algorithm for a prediction filter which is of a model free [11] for exploiting temporal correlation. The data dynamics of the projected methodology are highly adaptive and computationally light weight. In PRESTO [12], a model is constructed by high tier proxies which helps in capturing the correlations in the observed data at each sensor having low tier. The model's observed data gets deviated from the predicted values when the remote sensors start checking the sensed data against this projected model and starts push data and captures anomalous trends. In PRESTO, consideration of only temporal correlation is done, but the nearby sensor spatial correlation is ignored. Various data gathering methods [13] have been projected the regulation of spatial characteristics active sensors In [14], to sample the data from different sources a linear model is projected by capturing the spatial correlation. Most sensor nodes can be put into sleeping mode with the help of this model, and the inference of their reading can be done with definite accuracy with the help of the linear combination of data set from working sensor nodes. However, in real time, numerous systems may not be linear. Furthermore, the methodology for choosing the correct working nodes has not been discussed in [14]. ASAP [15] creation of sub clusters is done with the help of correlated sensor nodes, a subset of samplers is selected from the sub cluster through which data is collected continuously. Spatial correlation helps in the prediction of the non-sampler data set. ASAP uses probabilistic models are used in ASAP for the validation of forced sampling periods only, thus the predicted data's error bound is not guaranteed. Anomalous trends between forced samples may not be noticed if the prediction of errors is not done properly. Based on the closely correlated sensor nodes a sub cluster can be constructed for the projected work. For every data collection round the spatial correlation is verified. The concentration is mainly done on either temporal correlation or spatial correlation in all the methods given above.

Some of the data gathering methodologies utilize both temporal and spatial correlations between the sensor data for the reduction of the communication overhead. BBQ [16] at first it uses multivariate Gaussian joint distribution for capturing the correlations of sensor readings. To estimate the non sampled sensor readings, it samples a minimal amount of sensor data from a WSN and for this process it utilizes a Gaussian joint distribution model. However, an expensive long training phase and an entirely detailed data set of every solitary sensor node with long period sufficiency is required for these kind of models. Second, A continuous model update with periodical data gathered by each and every sensor and all the related information about the sensor nodes to ensure the perfections of this kind of models.

EEDC [17] similar nodes from different clusters are selected. The approximation of one sensor node from another is done within a cluster. Thus the sensor nodes are scheduled inside the cluster to work alternatively for the maintenance of energy. With the help of piecewise linear approximation temporal correlation can be executed, here the reduction of time series is in short line segments. Being a centralized methodology scalability issues have been undergone in EEDC. Secondly, there is no data error bound guarantee for non-sampled nodes in EEDC.

The projected work deals with the accumulation of the goals of an ideal data gathering system. The distributed system work is applicable to WSN having any magnitude. Considering solitary node which has the construction of similarity based clusters through which the systems spatial correlation is exploited. Dual prediction based reporting exploits temporal correlations. To achieve spatio-temporal correlation based on the reduction of data the given two approaches are exactly combined into collective prediction. Light weight algorithms are used in this system, which are more likely suitable for constrained resources. Each and every change in the data can be easily adapted by this system in both domains namely spatial and temporal. The system does not exceed the error bound and maintains the collected data within the allocated error bound.

III. COLLECTIVE DUAL PREDICTION

3.1. System Model &Overview

The portrayal and indication of rudimentary functionalities are described in an overview SAMI architecture and a transitory explanation on the set of mechanisms employed. By exploiting the spatial and temporal correlation between the sensors data in the projected work there is a reduction of data communication in the network. The sub clusters starts grouping the highly correlated sensor data where the inferences of spatial correlation between the sensor data are processed. A solitary sampler node is a representation of a sub cluster, suppressing the neighboring nodes redundant data. LMS filter estimates the sensor data's temporal correlation for the prediction of future data. Since we have the predicted data, a particular subset of data which deviates from the desired data is transmitted. This method helps in filtering the entire spatial and temporal redundant data. By introducing the collective prediction methodology the identification of anomalous trends from sampler nodes become simpler and those anomalies are communicated to the sink.
Three layer constructions are followed in this projected system. There are N numbers of nodes in the bottom layer which are haphazardly distributed over the field. Each node itself behaves as a system, which can compute, senses where finite energy source powers the communication modules. Group of spatially nearer nodes, which are associated with high energy cluster head are embedded in a node cluster, which behaves as the second layer. The data of the node get transmitted to the CH, which aggregates and forwards the data to the base station. Clusters build the third layer, where the closely correlated nodes are partitioned into numerous sub clusters. Each sub cluster has a sub cluster head (SCH) to represent it. The SCH estimates the data generated by SCH for the representation of the entire sub cluster. Fig.1 shows the three layer architecture.

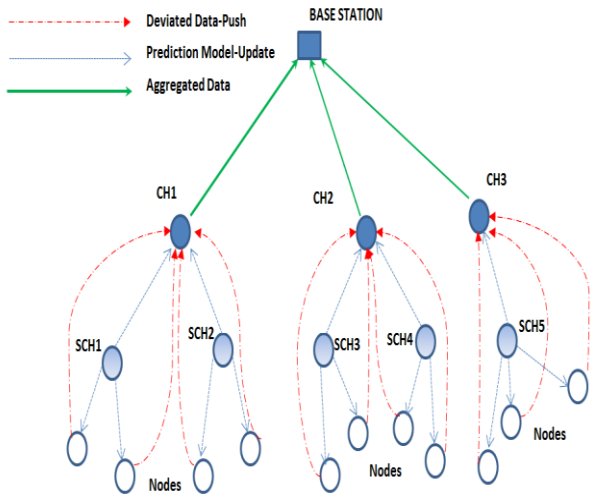

Fig.1. Three Layer Architecture of SAMI

The representation of the workflow is done in four phases. Using weight based passive clustering method the energetic nodes are designated as cluster heads and the clusters are fabricated around them. Based on their data similarity, the CH starts gathering data from its members for further portioning them to sub clusters. Representation of sub cluster (SCH) is made by the node which is having utmost energy. A temporal correlation model is constructed by SCH using an LMS based filter with the observation of its previous model. In this model the prediction of future data becomes simpler with the help of user defined error forbearance. The sub cluster members and CH share this model. Whenever there is a deviation between the predicted data and observed data yonder the temporal error threshold, every haphazard sample starts comparing the predicted data and observed data and updates the model, this operation takes place in fourth place. At every immediate sample, when there is a larger deviation in the error threshold, then the sub cluster members start comparing

the predicted data with its very own observations and interconnect the data. At every sampling instant, the CH twitches predicting the data based on the design and checks for any updates available from SCH and any interconnects from the sub cluster members.

Turan's theorem based spatial correlation by means of *Extremal Graph Theory*
Regulates the largest size, configuration explained in [18] with the help of the property given below
Given a fixed graph,

$$ex(v, R) = \max \{ t(E) | R \not\subset E, |S(E)| = v \}$$

When there is no update available from SCH then the model is perfect for that time instant. The entire sub cluster data are considered for approximation. The sub clusters inspect for any communication from other members, if so the equivalent data gets replaced for the corresponding sub cluster's data series.

3.2. Energy efficient passive clustering

Most of the WSNs achieve scalability by clustering physically nearby nodes and also by checking the route comforts and bandwidth maintenances. By undertaking various complex inferences, the reckoning load is distributed by the cluster at the CH. Since there are numerous computations to be performed in our algorithm with various data series, we partition the network into a collection of spatially nearer nodes which are headed by an energetic node. In this algorithm for the CH election, we follow a deterministic methodology which guarantees constant distribution of CHs. In the projected work, nodes with high energy are elected as CH to achieve the desired energy efficiency. During the clustering process to reduce the consumption of energy the method of passive clustering has been projected, where the delay of proclamation is defined as the function of node's remaining energy. As CHs are burdened with multiple tasks, the rate of energy depletion increases. Hence the CH elected should have an extraordinary remaining energy. Thus the projected work uses the remaining energy as the weighted parameter for CH election. Obtaining remaining energy is again an interior task which does not need any communication.

Selection of Cluster head in passive clustering depends upon "first declaration wins" rule, consequently the first declared node becomes the CH. In the former works, the declaration delay is haphazard. Haphazard delay may result in selection of nodes with less energy as in the case of CH, therefore the system's energy efficiency decreases. In the projected work, to elect most suitable nodes, the declaration delay is made inversely proportional to the node's remaining energy. Once the declaration delay terminates, the node declares itself as CH. If a node receives a declaration before the termination of its own declaration delay, it ceases itself from the contention to become CH. The waiting time T_w of node n is given as

$$T_w(n) = K/E_{res} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, E_{res} is the remaining energy of node 'n',
k is a constant;
If a node receives multiple declarations from different nodes, it selects the nearest node as its CH and associates with it, by this process clusters are formed.

3.3. Exploiting Spatial correlation

In a meticulously settled WSN due to spatial proximity closer nodes senses similar data, without any informational value substantial amount of network energy is consumed while sending spatially similar data over the network. When the network is conventionally clustered, the cluster heads receive all the nodes in the cluster, where the data are combined and sent to the sink. In a clustered combination scheme, cluster filters the spatially redundant data, accordingly discards the additional flow of insignificant data. However, conquering the redundant data at the node itself might be a better option than filtering it out at the CH. Here in the projected work, based on the sensor time series, depending upon the data similarity different sub clusters are assigned by the cluster head. A SCH node each time reports a data to CH for each corresponding sub cluster, accordingly the spatially redundant data are filtered indigenously.

The formation of sub cluster is done in three creases. Predominantly, a node with high energy is identified as SCH. Next, its closest neighbors are identified. Next, the SCH's data series are compared to other neighbors. The magnitude and trend similarity are compared. The sub clusters which heads SCH adds the nodes when the magnitude and trend of the neighbors are similar. With the consideration of other cluster members the node with secondary level energy is identified replications are made in this process until all the cluster nodes are added to a sub cluster.

3.3.1. Sub Clustering

Through the SCH, the data in each sub cluster are reported to CH. Accordingly the other nodes should have lower residual energy when compared with SCH. All the cluster members residual energy is estimated by the cluster head. Then high energy node is selected as SCH by CH. Once the identification of SCH node in the cluster is completed, then the nodes with the Distance D_{th} can be listed with the help

of CH. where D_{th} represents the maximum threshold distance for each sub cluster. The pair of nodes is said to be spatially similar, when the distance D is less than D_{th} .Implication of data similarity of nearby nodes and SCH is done. The documentation of similarity is in two creases, first the similarity between the magnitudes and trend are measured.

Let x (x_1, x_2,…, Xn) represents the time series of the node x and y represent the time series of the node y (y_1, y_2,…, y_n). The Euclidean distance between two time series is represented by CH which also denotes the magnitude similarity between the time series. d (x, y) is the Euclidean distance between two time series and it is denoted as follows

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

The linear affiliation between two time series can be dignified by the correlation coefficient. Pearson's linear coefficient processes the correlation coefficient between two time series with the respect to CH. The equation given below denotes the Pearson's correlation coefficient,

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

α similar data series are represented here, if T (x, y) > 0.9 and d (x, y) < α. If the above conditions occur, then the node y is added to node x's represented sub cluster by the CH. Sub cluster's formation is done in such a way that the members in the sub cluster are α similar with a representative for each sub cluster. While all the α of the representative data are identical as the CH communicates only with the representative data, the remaining data's error bound can be approximated by α.With the help of this scheme significant reduction in data communications can be made within the clusters As a result the sensor's complete energy expense reduces.

3.4. Exploiting Temporal correlation

The node successive similarity depends upon the temporal correlation of a time instant. Over a certain time span a major amount of redundant data gets added to this temporal correlation. To identify a subset of sensor readings and to reduce the consumed energy, within a defined accuracy the exploitation of temporal correlation among the data is done. The base station avoids the prediction of already delivered data, so as to reduce the communication. The sensor data interpretation is done in the time domain for the digital filters, therefore the prediction of future values becomes simple. The inference of the short term linearity of the signal can be done with the help of LMS filter. A linear combination of future data can be predicted with the help of the latest data history, based on the previous inference.

Using prediction based reporting the data can be reduced in the structure of clustered data aggression At sensor node and CH the identical prediction filters are defined. The actual detected value in the sensor node is compared to the filter value predicted, at each sampling instant t. No data gets transmitted when the difference between the CH and sensor node are lesser than that of the threshold. The data gets transmitted to the CH when the difference is higher than the threshold value. Accordingly, only a portion of data is transmitted.

With three modes the prediction based reporting is executed. At each sampling instant t, the sensor data transmit the data to CH, instantaneously the prediction engine updates the corresponding coefficients towards convergence based on the deviation. This mode is defined as initialization mode. If the error threshold of β is greater than that of the deviation, then there is a convergence in the prediction, for M successive predictions. The filter switches to standalone mode so that the filter model communicates with the CH.

Turan's theorem based temporal correlation to overcome large subset problems

$$C_b \leq \alpha \geq C_p$$

Equality of the spans can be computed using Turan's graph as discussed in [19]

$$\frac{k-1}{2k}(2 \propto -1)p^2$$

There is a comparison between the actual sensed value and the filters predicted value in standalone mode, at each sampling instant t. The filter model is considered to be accurate for the time instant t, when there is a small deviation in the actual threshold value. The data transmission does not occur in this case. Meanwhile the model is well acknowledged by CH, the predicted value can be computed and can be considered as actual observation's time instant approximation. As long as the model's observed value is accurately predicted, communication is not required between the sensor node and CH. This mode switches to normal mode when the error exceeds β. The data

gets transmitted to CH during the normal mode. To converge the prediction with the desired value, the weight of the prediction filter gets adjusted. The filter model gets updates to CH and again switches to standalone mode when the prediction is converged.

3.4.2. nLMS based Prediction filter

In the projected work the nLMS algorithm builds the prediction filter. Brief explanations of the functional features of an nLMS based prediction filter are discussed below. At an instant k, a data stream of x is sampled by a linear adaptive filter with respective length n, the notation is given as x [k] and the predictions are calculated as y [k] = w^T [k]· x [k], which is an effective linear combination of earlier n data stream samples, balanced by the corresponding weight vector w [k]. The corresponding signal d [k] and the output y [k] are compared.

$y[k]=\underline{w}^T[k]\underline{x}[k]$	Filter output
$e[k]=d[k]-y[k]$	Inference Error
$\underline{w}[k+1]=\underline{w}[k]+\mu\underline{x}[k]e[k]/\underline{x}^T[k]\underline{x}[k]$	Weights adaptation
$\mu = (1 / E_x) / D$	Step size calculation
$E_x = \frac{1}{M}\sum_{k=1}^{M} x[k]^2$	Mean input power

Table.1.LMS model

3.4.3 VSS-nLMS prediction

The prediction error e [k] can be computed as: e [k] = y [k] −d [k] and fed into the adaptation algorithm. In order to minimize the mean square error, at each time step k the filter weights are updated. At every time step, the step size is normalized in the normalized LMS filter, results in the reduction of the input sensitivity. The table describes the functional model of the nLMS algorithm.

Step size plays a vital role in the achievement of energy efficiency and data accuracy in the prediction report. Hence the step size is content dependent, it does not have any specific optimum value. When the deviation is high, larger step size conquers faster merging. Smaller step sizes realizes steady state prediction near the point of convergence. Prediction model helps in adapting the step size by achieving a substantial speed of convergence and reduce in the deviation. During different states of prediction to control the step size we introduce an integer D.

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

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

3.5. Exploiting spatial, temporal correlation (Collective Prediction)

With the help of independent prediction filters, CH receives data from all the CH members in the methodology of conventional prediction based reporting. Highly correlated data source located in close vicinity characteristics a sub cluster and a solitary SCH node represents it in this projected work. Accordingly the SCH node predicted by the prediction filter is considered sufficient for the entire sub cluster. Using a common model, comparison between SCH and CH are analyzed based on the prediction based reporting. As there is a close correlation between the neighboring representative nodes, the data which have been generated by the representative node with respect to magnitude and trend is the resemblance of the entire sub cluster. The collection of representative characterizes the data approximation of the entire sub cluster whose error bound falls within α. There is no guarantee of error bound for the non-sampler nodes. Here in this work, we have projected a novel collective dual prediction methodology, which helps in the detection of spatial irregularities during the data collection and also rectifies it with the help of model driven push scheme. On behalf of a sub cluster, prediction model is

constructed by each node based on the data observed by it. The past observation uses correlation to forecast the value prospective to be seen on subsequent instant t. The CH and the other sub cluster members receives the model and its parameters.

The model is executed by the sub cluster as follows: the value predicted by the model is compared with the actual sensed value at each sampling instant t. The model is said to be α similar for that instantaneous period when the threshold is greater than the value predicted by the model and the actual sensed data. None of the sub cluster members transmit data in this case. The values can be predicted and used as an α approximation of the actual observation of the particular sensor node as the CH knows the model. The communication between the sub cluster member and CH is not required as long as the model predictions and observed values are similar. In divergence, if the comparison between the model predicted data and the sensed data exceeds a threshold, the sensed value is sent to the CH. Thus the sub cluster member's sends data only when there is a deviation in the value of the predicted value of the common model, by the means of capturing the trends deviation. Such a collective dual prediction methodology reduces communication overhead by reducing the SCH per sub cluster to one, while guaranteeing that deviating patterns of the other node's data are never missed. The projected system occupies VSS-nLMS based prediction filter for constructing prediction models, which is computationally inexpensive and provides optimum level of accuracy. Since trends in sensed values may change by that time, a model constructed using historical data may not replicate the current trend continues. The representative node appraises the adaptive filter model parameters during trend changes, so that the model can continue to reflect current observed trends. Our methodology integrates active feedback between the CH, representative and other sub cluster members' results in high reliability of data with considerable energy maintenance.

IV. Experimental Classification Results and Analysis

The actual purpose of implementing SAMI is to accomplish an energy efficient data collection. The energy efficiency is measured in terms of compact number of communication packets. The methodology achieves energy, maintenance at the cost of marginal tolerance in the data accuracy, hence it is vital to explore the average error of data acquired at the base station in correspondence to the original data observed at the sensors. There are abundant measures for approximating 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 projected work SAMI is appraised on MATLAB platform. In order to inspect the performance of SAMIwith large-scale networks, large traces of spatially correlated data set are generated significant to the mathematical model explored in [20] [21] [22], through which the model parameters are extracted from small-scale real data sets [23]. The work is appraised by comparing the performance with other energy efficient data gathering methodologies. Then the influence of spatial and temporal error threshold on the performance is appraised. The influence of the cluster size affects the system performance and scalability.

4.1 Comparison with other methodologies

The work is appraised with the amount of data reduction, for different error thresholds(Δ). With the increased Δ, the communication cost is concentrated. The projected work SAMI is equated with other data reduction methodologies like PRESTO and ASAP. PRESTO methodology involves model driven push, where the temporal correlation of the sensor data is utilized for data reduction. In ASAP, the spatial correlation sandwiched 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 measured as a spatial error threshold, in PRESTO, the error threshold is measured as a temporal error threshold. In SAMI,the error threshold (Δ) is split into two equal error thresholds namely temporal error threshold (α) and spatial error threshold (β). ($\Delta=\alpha+\beta$). The data reduction accomplished by SAMI overtakes both PRESTO and ASAP, since it jointly exploits spatio-temporal correlation among the sensor data.

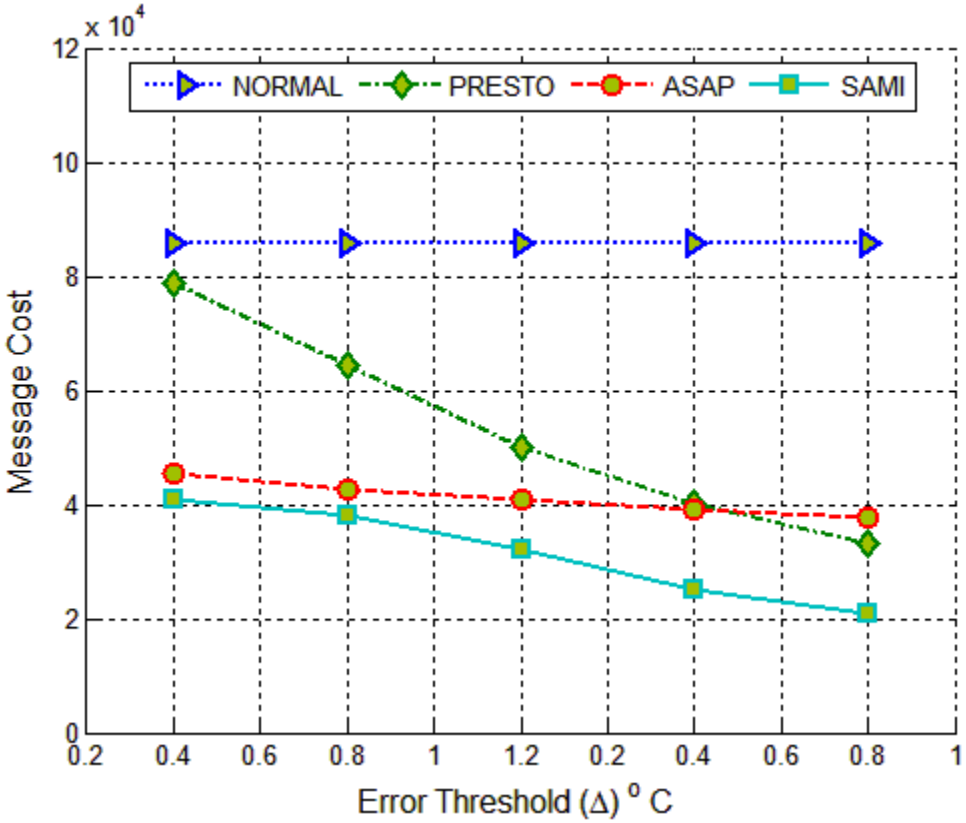

Fig.3. Message cost Vs Error Threshold (CODE, ASAP and PRESTO)

In SAMI, the data reduction is in two creases. 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 SAMI achieve inferior message cost than PRESTO.SinceSAMIdiverges the error threshold, during low Δ values, spatial and temporal models knowledges tight constraints.This obliges frequent updates to ensure the data within the specified error threshold.At Δ=0.2°C, SAMIcan condense only 40% data. When Δ increases, the temporal data reduction is steep along with reasonable reduction in active reporting nodes. Thus the dual reduction methodology performs well and SAMI overtakes both PRESTO and ASAP. At Δ=1°C, SAMI can reduce about 75% data. The message comparison is shown in Fig.3.

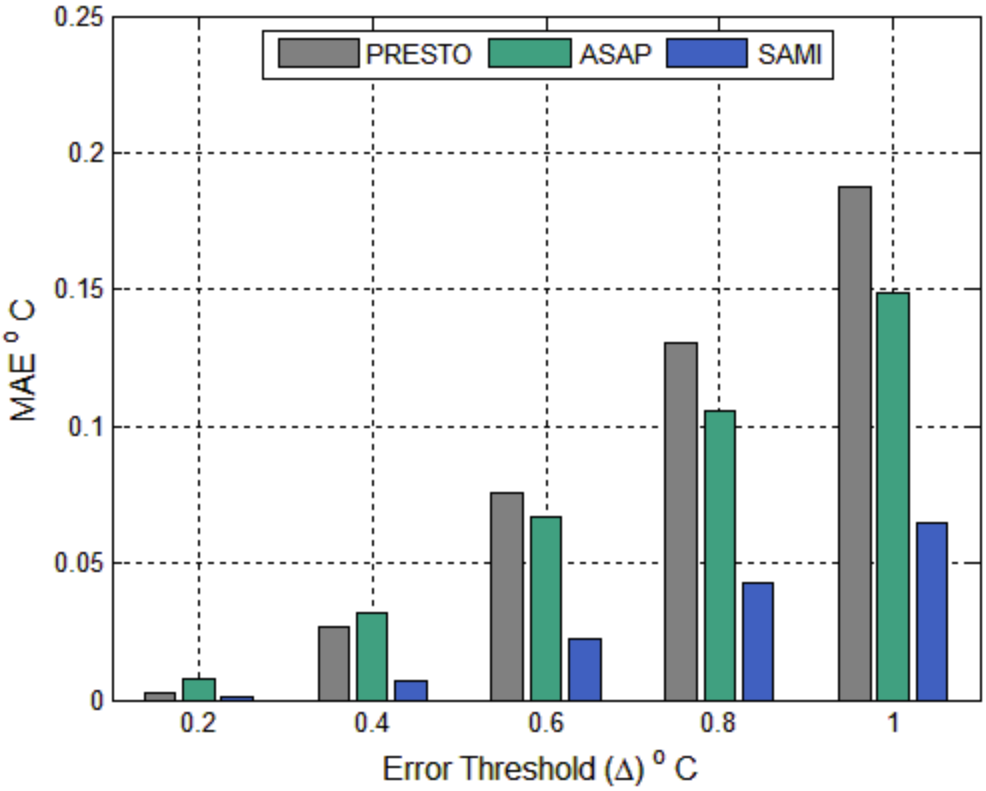

Fig.4. MAE Vs Error Threshold (CODE, ASAP and PRESTO)

Another metric to appraise the performance of a data gathering methodology is to analyze the mean absolute deviation between the observed data and the data collected at the base station. The efficiency of the system towards sinking the deviations is measured using this performance. In the projected 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 glibber for SAMI and increase brusquely for the ASAP and the PRESTO. Compared to PRESTO and ASAP, the mean deviation is much lesser in SAMI, due to the divergence of error threshold into spatial and temporal error thresholds and the coinciding of spatial and temporal errors. As in Fig.4. For higher Δ values, the PRESTO has higher mean deviations, since high temporal error threshold allows the prediction to diverge for larger values. At Δ=1°C, the deviation of SAMI is less than a half of the ASAP and PRESTO.

4.2 Impact of Spatio-Temporal Error Threshold

Since the system comprises both spatial and temporal data reduction, we analyze the influence ofα and β individually on the performance of the projected system. The temporal error threshold αresolves the frequency of trend change updates. Greater the α, lesser the frequency of updates and vice versa. The β decide the volume of active nodes to report the trend changes in the network.The low value of β upsurges the number of active nodes, which upsurges the spatial granularity of the data observation. The higher value of β reduces the active reporting nodes, thus preserves substantial amount of energy. The right combination of α and β better employs off between the data accuracy and energy maintenance. The α value is incremented from0.2°C to 0.8°C and for each α value, β diverses from 0.2°C to 0.8°C. For every blend of α and β, the message costs and mean data deviations are estimated. From the Fig.5., it is detected that the influence of spatial error threshold on data reduction is smooth, but the impact of temporal error threshold of data reduction is sharp. When we rise the spatial error tolerance the reduction of vigorous nodes is limited by distance threshold and the maximum number of nodes in the cluster. Consequently, increased spatial error tolerance cannot further reduce the number of vigorous nodes.

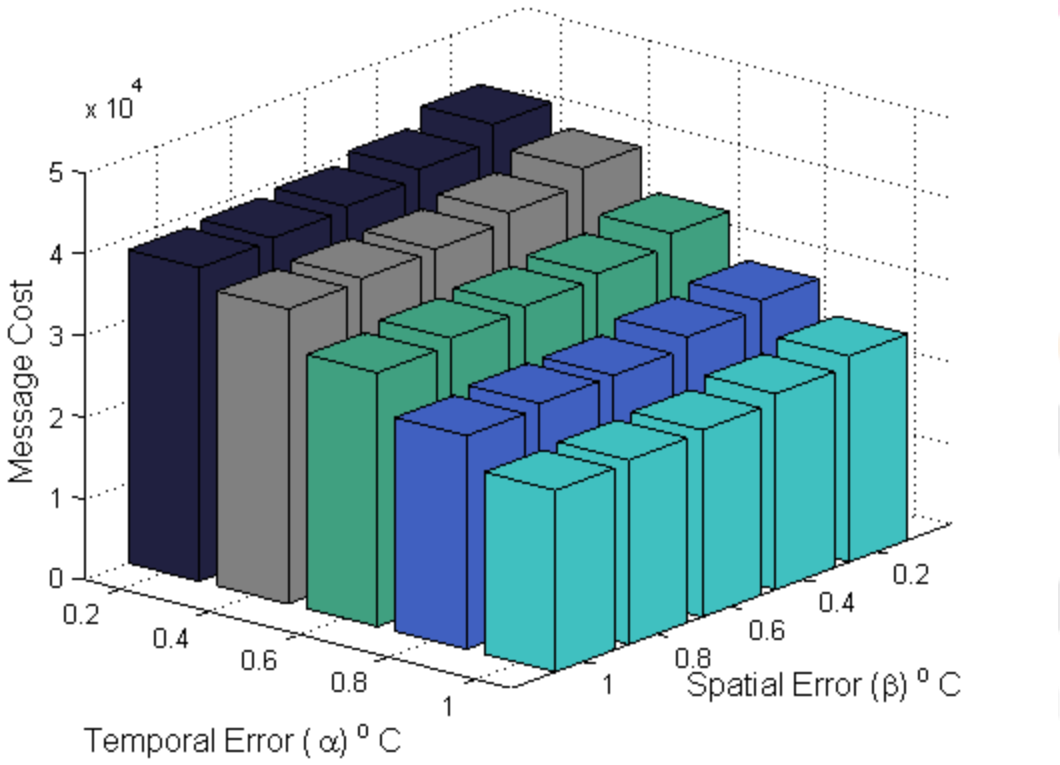

Fig.5. Message Cost Vs Error Threshold (various Spatio-Temporal Error)

The system syndicates the compensations of both spatial correlation and temporal correlation among sensor data,hence it is crucial to appraise separable performance on message cost to appraise their importance.Here we demonstrate the discrete message costs of temporal correlation based data collection and spatial data deviance updates for unlike combinations of α and β. The temporal correlation based data reporting is between the illustrative node and the CH. The spatial data modifications are among the sub cluster members and the CH. This appraisal helps in categorizing the right combination of α and β to accomplish an effectual data collection.from Fig.6, it is perceived that larger temporal and smaller spatial error threshold is the judicious choice to achieve significant data reduction laterally with optimal accuracy on collected data.

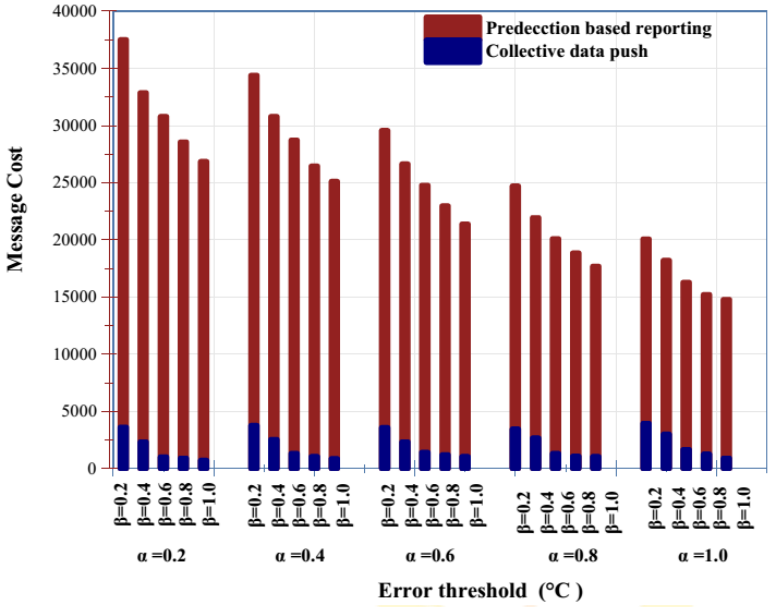

Fig.6.Message Cost Vs Error Threshold (show temporal data and spatial data)

In the identical way, the effects of α and β are analyzed on the mean absolute error of the data collected.When there is increase inα, the MAE also increases.The spatial observation error increases with the increase in 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 progressively with increased β as shown in Fig.7.

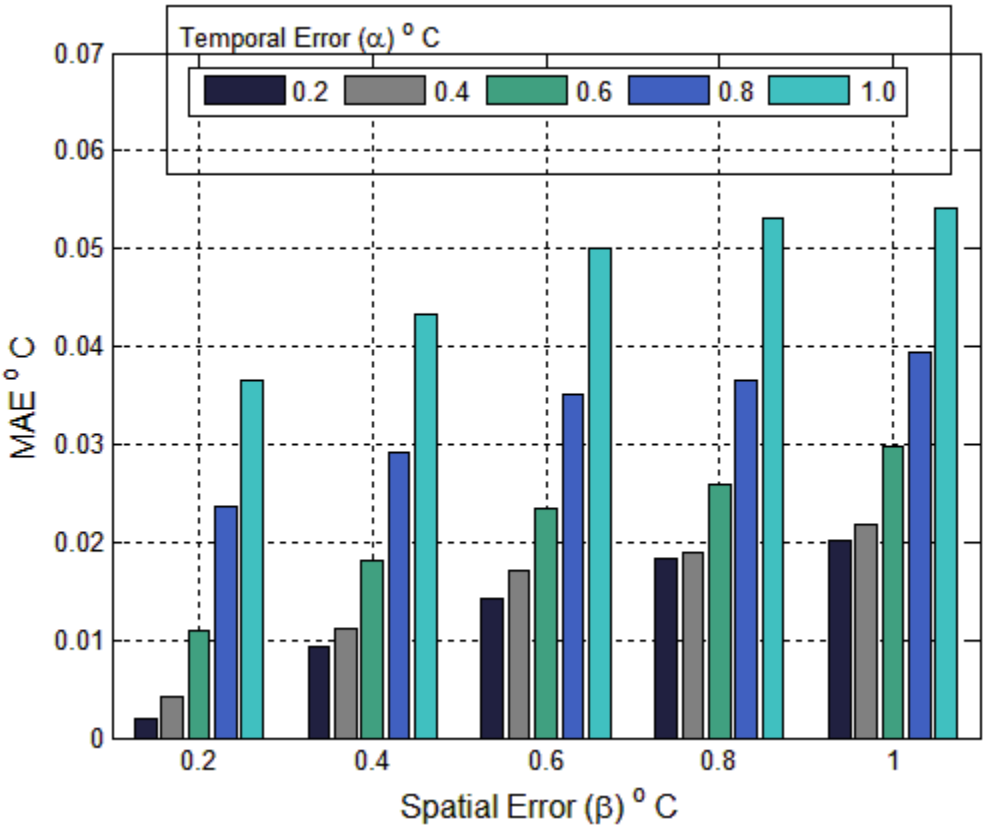

Fig.7. MAE Vs Error Threshold (various Spatio-temporal Error)

4.3 Impact of Cluster size

Here we explore the influence of cluster size on the efficiency of our projected system. The number of nodes in the cluster is certain with the respect to the cluster size. The larger cluster size fetches in more nodes, therefore the nodes per sub cluster also increases. The heavy weight sub clusters condenses the number of active reporting nodes, accordingly increases the energy efficiency.The data reduction and

accuracy of the framework with different cluster sizes are indicated in Fig.8.The augmented cluster size helps in achieving more spatial reduction. In a small cluster size, the spatial data reduction with respect to augmented 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 precipitous 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 augmented cluster size brings a double benefit for the projected data reduction methodology.Since the number active nodes are less, the total trend updates are also condensed. This further reduces the nodal communications. From the simulations, it is perceived that the cluster size has no direct impact on the temporal data reduction, but has high influence on the spatial data reduction. The mean data deviation is also analyzed for contrasting cluster sizes. Augmented spatial error threshold increases the data reduction in the cost of increased data deviation. In small clusters, the difference in data deviation with respect to increased spatial error threshold is negligible. In huge clusters, when the spatial threshold increases, more nodes are put into passive mode. Therefore the data deviation also increases snappishly.

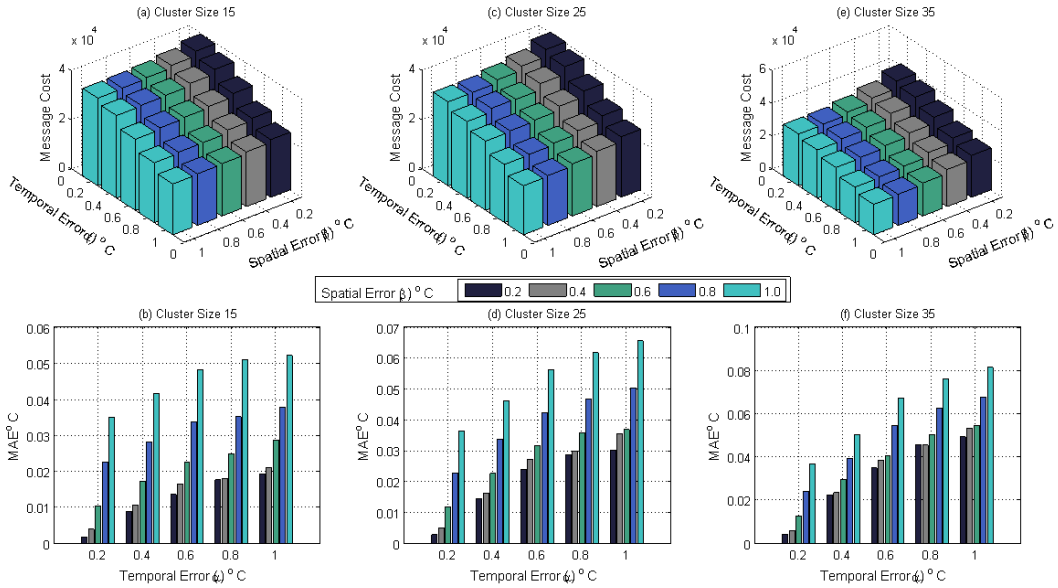

Fig.8. Impact of cluster size

4.4Scalability

In a distributed data gathering methodology, the scalability is a significant parameter. The projected work is appraised on networks of numerous scales. The performance improves the size of the network. Increased number of nodes increases the node compactness of the network. When node compactness increases more nodes get into close vicinity, results in a substantial increase in spatially correlated data.. This close vicinity increases the size of sub clusters. Fig. 9. shows the increased size of the network exponentially increases the number of sub clusters in the network, hence the percentage of active nodes are reduced.

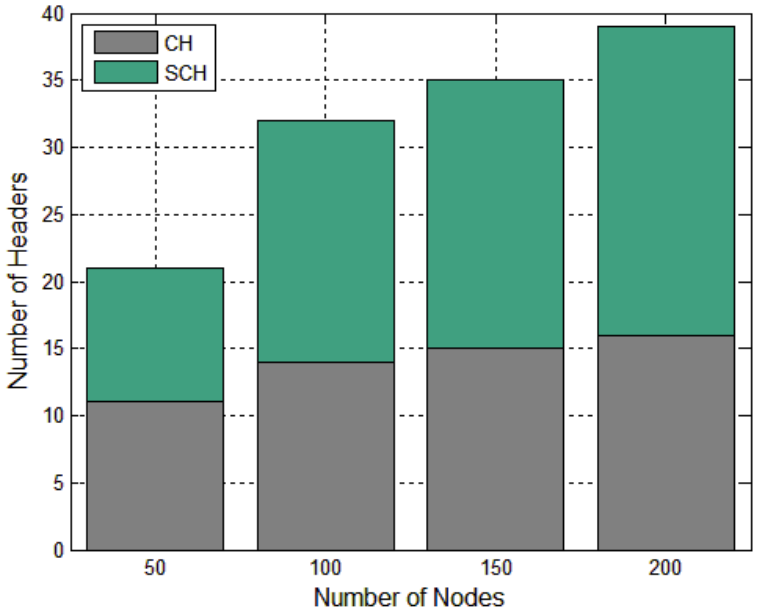

Fig.9. Cluster and Sub Cluster (various network scales)

The Fig.10. Shows the projected work which outfits well with the large scale networks. When there is an increase in the number of nodes, the data load of the network increases.Here in the projected 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..10 . It is incidental that with the increase in node compactness, SAMI increases the percentage data reduction from 75% to 87%.

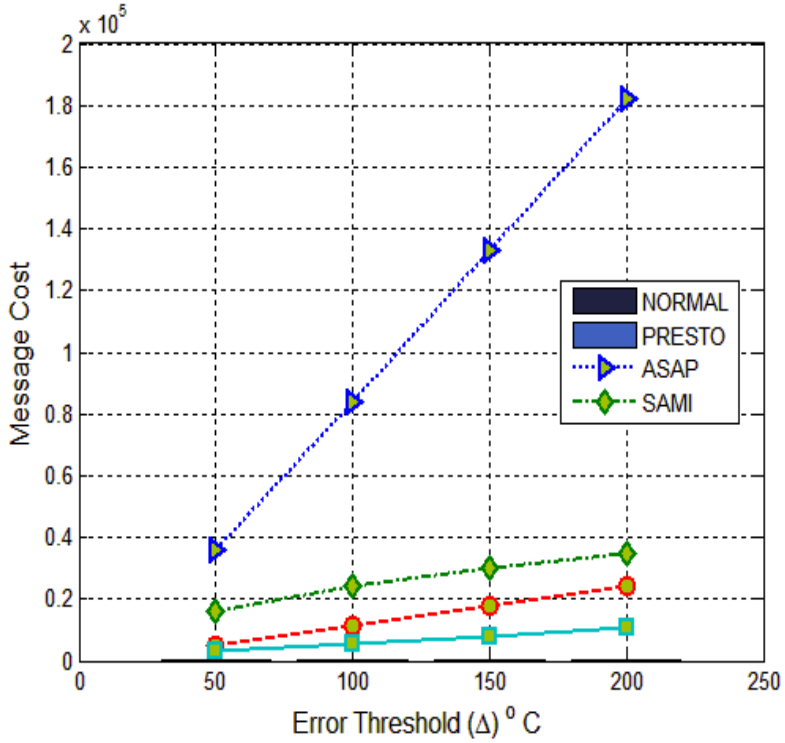

Fig.10. Message Cost Vs Error Threshold (various network scales)

V. CONCLUSION

SAMI accomplishes less deviation in the accuracy of the data collected and achievement in two level data reduction The work has been appraised by the amount of data reduction and mean absolute data deviation. Subsequently the work results in the combined reduction of the spatial and temporal redundant data . The data reduction has improved to multiple folds than the previous works. It is highly a directed mechanism which assures user specified error threshold in spatial and temporal aspects. The increase in the system

performance increases the error tolerance. The system has proven to be exceedingly scalable. The influence of various parameters are detaily analyzed. There is a 75 % reduction in data transmission with 0.07°C mean absolute deviation. The future work implicates vigorously adjusting the spatial and temporal error thresholds based on data dynamics and spatial discrepancies.

VI. REFERENCES

[1] M. Li and Y. Liu, "Underground Coal Mine Monitoring with Wireless Sensor Networks, "ACM Trans. Sensor Networks, vol. 5, no. 2, pp. 1-29, 2009

[2] M. Li, Y. Liu, and L. Chen, "Non-Threshold Based Event Detection for 3D Environment Monitoring in Sensor Networks," IEEE Trans. Knowledge and Data Eng.,vol. 20, no. 12, pp. 1699-1711, Dec. 2008.

[3] G Li, J He, Y Fu, Group-based intrusion detection system in wireless sensor networks. Comput. Commun.31(18), 4324–4332 (2008)

[4] M. Li and Y. Liu, "Underground Coal Mine Monitoring with Wireless Sensor Networks, "ACM Trans. Sensor Networks, vol. 5, no. 2, pp. 1-29, 2009

[5] D. Culler, D. Estrin, and M. Srivastava, "Overview of sensor networks," Computer, vol. 37, pp. 41–49, 2004.

[6] D. Chu, A. Deshpande, J.M. Hellerstein, and W. Hong, "Approximate Data Collection in Sensor Networks Using Probabilistic Models,"Proc. 22nd Int'l Conf. Data Eng. (ICDE '06),2006

[7] Anastasi, G., Conti, M., Di Francesco, M., and Passarella, A. 2009. Energy conservation in wireless sensor networks: A survey. Ad Hoc Networks 7, 3 (May), 537–568.

[8] S. Chatterjea and P. Havinga, "An Adaptive and Autonomous Sensor Sampling Frequency Control Scheme for Energy-Efficient Data Acquisition in Wireless Sensor Networks," Proc. IEEE Fourth Int'l Conf. Distributed Computing in Sensor Systems (DCOSS '08), 2008

[9] Carlos.Carvalho, Danielo. G.Gomes, Nazim Agoulmine and José Neuman de Souza "Improving Prediction Accuracy for WSN Data Reduction by Applying Multivariate Spatio Temporal Correlation," Sensors , vol. 11, pp. 10010–10037, Oct.2011.

[10] Li and Wang " Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks" EURASIP Journal on Wireless Communications and Networking,2013,2013:85.

[11] Lazaridis I, MehrotraS," Capturing sensor-generated time series with quality Guarantees," *in Conf.* 19th international conference on data engineering,2003, pp.429-440.

[12] Santini S and Römer K, "An adaptive strategy for quality based data reduction in wireless sensor networks*," in Conf. networked sensing systems*, 2006, pp. 29-36.

[13] Ming Li ; Ganesan, D. ; Shenoy, P." PRESTO: Feedback-Driven Data Management in Sensor Networks", IEEE/ACM Transactions on Networking (Volume:17 , Issue: 4), Aug. 2009,pp.1256 - 1269

[14] L Villas, A Boukerche, H Oliveira, R Araujo, A Loureiro, A spatial correlation aware algorithm to perform efficient data collection in wireless sensor networks. Ad Hoc Netw.11(3), 966–983 (2013)

[15] F. Emekci, S.E. Tuna, D. Agrawal, and A.E. Abbadi, "BINOCULAR: A System Monitoring Framework,"Proc. First Workshop Data Management for Sensor Networks (DMSN '04),Aug. 2004.

[16] Gedik, B. ,Ling Liu ,Yu, P.S." ASAP: An Adaptive Sampling Approach to Data Collection in Sensor Networks" IEEE Transactions on Parallel and Distributed Systems, (Volume:18 , Issue: 12) Dec. 2007,pp- 1766 – 1783.

[17] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W.Hong, "Model-Driven Data Acquisition in Sensor Networks," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB '04),2004.

[18] Tolhuizen, L.M.G.M. , "The generalized Gilbert-Varshamov bound is implied by Turan's theorem," IEEE Transactions on Information Theory, Vol.43 , No.5, pp. 1605-1606, Sep 1997.

[19] Reibiger, A. "Networks, multipoles and multiports", Circuit Theory and Design (ECCTD), 2013 European Conference on, On page(s): 1 - 37

[20] Kasami T, Shu Lin , Wei VK, Yamamura Saburo, "Graph theoretic approaches to the code construction for the two-user multiple- access binary adder," IEEE Transactions on Information Theory,Vol.29 , No. 1 ,pp.114 - 130,Jan 1983.

[21] Chong Liu,Kuiwu, 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, July 2007, pp-1010-1023

[22] A. Jindal and K. Psounis, "Modeling Spatially-Correlated Sensor Network Data," ACM Trans. Sensor Networks, vol. 2, no. 4, pp. 466-499, 2006

[23] [Online]. Available:.http://www.intelresearch.net/berkeley/ index.asp