Energy Aware Node Selection for Cluster-based Data Accuracy Estimation in Wireless Sensor Networks

Jyotirmoy Karjee
Centre for Electronics Design and Technology
Indian Institute of Science, Bangalore, India
Email: kjyotirmoy@cedt.iisc.ernet.in

H.S Jamadagni
Centre for Electronics Design and Technology
Indian Institute of Science, Bangalore, India
Email: hsjam@cedt.iisc.ernet.in

---ABSTRACT---

The main objective of this paper is to reduce the number of sensor nodes by estimating a trade off between data accuracy and energy consumption for selecting nodes in probabilistic approach in a distributed network. *Design Procedure/Approach:* **Observed data are highly correlated among sensor nodes in the spatial domain due to deployment of high density of sensor nodes. These sensor nodes form non-overlapping distributed clusters due to high data correlation among them. We develop a probabilistic model for each distributed cluster to perform data accuracy and energy consumption model in the network. Finally we find a trade off between data accuracy and energy consumption model to select an optimal number of sensor nodes in each distributed cluster. We also compare the performance for our data accuracy estimation model with information accuracy model for each distributed cluster in the network.** *Practical Implementation:* **Measuring temperature in physical environment and measuring moisture content in agricultural field.** *Inventive /Novel Idea:* **Optimal node selection in probabilistic approach using the trade of between data accuracy and energy consumption in a cluster-based distributed network.**

Keywords - **Spatial correlation; distributed clusters; data accuracy; energy consumption; tradeoff; wireless sensor networks.**

Date of Submission: December 25, 2011   Date of Acceptance: January 26, 2012
----

1. Introduction

Recent improvements in wireless communication have made a drastic development over wireless sensor networks and wireless embedded systems. Because of the reliable cost, ease of development, small size, wireless sensors are used in many applications such as collecting data of any event like temperature, humidity, seismic event etc. Wireless sensor is a small processing device called node which captures raw data from the physical environment[1], process it, communicate the raw data wirelessly among other nodes and finally transmit the collected raw data to the sink node. The raw data collected by the sensor nodes are generally spatially correlated [2] among other sensor nodes in the physical environment. As the density of the sensor nodes increases the spatially proximal sensor observation among the sensor nodes are highly correlated [3] in the sensor field. The sensor nodes form distributed clusters [4] in the sensor field due to high data correlation among the sensor nodes. Formation of distributed clusters minimizes data collection cost [5]. LEACH [6] gives a

clear idea about the formation of distributed dynamic clusters in the sensor field according to priori probability. Each dynamic distributed clusters have a cluster head (CH) [7,28] node which collects the observed data and transmits it to the sink node. Like LEACH, SEP [8] forms distributed clusters in the heterogeneous networks. According to literature [9], sensor nodes forms spatial correlation of data among them where the data correlation shows Gaussian distribution with zero mean. In literature [10], a grid based clustering method is proposed. But these types of theoretical clustering methods rarely happen in original scenario. In literature [11] a disk shaped circular clustering algorithm is proposed where sensor nodes groups into disjoint set with a designated CH node. This type of disk shape cluster doesn't form in real case. In practical scenario, clusters are irregular in shape and size. In literature [4], authors proposed a distributed clustering algorithm with different shape and size based upon shortest distance among associated sensor nodes and CH node in the sensor field. In literature [16], authors proposed a new non-overlapping distributed clustering algorithm with different sizes based upon spatially

correlated data among sensor nodes in the spatial domain. We adopt the clustering algorithm [16], since this non overlapping clustering algorithm has more practical sense of implementation in the spatial domain. The size of each distributed cluster in this clustering algorithm is based upon a threshold value given in data correlation model [4,16] in the sensor field.

According to literature [12,13,14,20], authors proposed information accuracy (distortion function) model where sink node can estimate the information accuracy for observed data sensed by all the sensor nodes. These proposed information accuracy model follows the single hop communication where the observed data are directly transmitted to the sink node. But in literature [4,15,16],authors proposed two hop communication where the observed data are transmitted to the sink node via CH node for each distributed cluster in the network. Generally observed data are directly transmitted to the CH node for each distributed cluster for aggregation [19,26,27] without verifying the data accuracy. This increases the redundant data in the CH node and also increases the transmission overhead in the network. Hence we verify the data accuracy using Minimum Mean Square Error (MMSE) estimator [25] before data aggregation at CH node for each distributed cluster in the network. Sometimes due to extreme external physical environment such as heavy rain fall, snow fall etc, some of the sensor nodes get malicious [18]. In such situation sensor nodes can sense and read inaccurate data in the cluster. In accurate data transmitted by malicious node may cause incorrect data aggregation at the CH node which also increases the data redundancy in the network. Hence it is very essential to estimate and verify the data before aggregation at the CH node for each distributed clusters in the sensor network.

In this paper, we use the proposed data accuracy model which uses MMSE estimator given in the literature [16] and compare with information accuracy [12] model .According to literature [12], MMSE estimation is done at each individual sensor node for the observed data before transmitting the estimated data at the CH node if we assume each distributed cluster in the network. Once the estimated data is received at the CH node transmitted by all the sensor nodes in the cluster, we average it and finally transmit it to the sink node. This increases the communication overhead in the network .But in data accuracy model, we calculate the MMSE estimation only at the CH node for all the sensor nodes in the cluster which increases the data accuracy for each distributed cluster and reduces the communication overhead in the network .

In literature [29], the author proposes a node selection scheme where the network reduces energy consumption by identifying redundant nodes with respect to sensing coverage. In literature [17], authors proposes selection of nodes using GB (Grid based) node choice algorithm which save energy in the network. Similarly in literature [20], authors minimizes energy consumption with optimizing the maximum information accuracy for a cluster based WSN by using an information accuracy aware jointly sensing nodes selection algorithm. Hence these work deals with

reducing the number of sensor nodes in the network with respect to sensing coverage and also finding an information accuracy aware jointly sensing nodes selection algorithm based on maximum information accuracy with minimum energy consumption. But in this paper we propose a novel probabilistic model to select an optimal sensor node in each distributed cluster by finding a trade off between data accuracy and energy consumption. We also find the probability by which the sensor nodes are in active mode and sleep mode in each distributed cluster subjected to data accuracy and energy consumption. Finally this leads to select an optimal sensor node which reduces the energy consumption and increases the lifetime of the network.

The main objective of this paper is to reduce the number of sensor nodes per cluster in the distributed network by finding a trade off between data accuracy and energy consumption in probabilistic approach. Hence the flow of this paper is given as follows. In Section. 2, we summarize our mathematical model proposed in literature [16] to extend our new work which is discussed in Section. 3. Initially we deploy all the sensor nodes in the sensor field. Sensor nodes form data correlation [16] among observed data sensed by them discussed in Section.2.1. Since the observed data are highly correlated among sensor nodes, sensor nodes form non-overlapping distributed clusters among them in the sensor field explained in Section.2.2. Once the sensor nodes form non-overlapping disjoint groups of distributed clusters, each distributed cluster perform data accuracy at the respective CH node discussed in Section.2.3. In Section.3, we propose a probabilistic model for selecting some of the sensor nodes to be switched on in each distributed cluster for transmitting the data to the sink node. In Section.3.1 and Section.3.2, we find by what probability the sensor nodes are switch on and switch off in each distributed cluster can be controlled subjected to data accuracy and energy consumption respectively in the whole network. This leads to save energy in the whole network by controlling some of the sensor nodes to be in sleep mode for each distributed cluster. It leads to find a trade off between data accuracy and energy consumption for each distributed clusters in the network discussed in Section.3.3. This trade off reduces node deployment cost and energy cost per cluster in the distributed network. Finally this trade off between data accuracy and energy consumption for each distributed cluster is implemented to reduce the number of sensor nodes for the whole network.

2. Cluster-based Data Accuracy Estimation in Spatial Domain

Initially sensor nodes are randomly deployed in the sensor field. If we keep on deploying the sensor nodes, the sensor node density increases and the spatial correlation of the observed data among the sensor nodes also increases in the sensor field. Since the spatial correlation of data increases among the sensor nodes, we get more accurate observed data from the sensor nodes. But as the number of sensor nodes increases, the node deployment cost and the energy

consumption in the network also increases. Hence we have to reduce the node deployment cost by reducing the number of sensor nodes such that we also achieve better accurate observed data from the sensor nodes with less energy consumption in the network.

In this section, we discuss a mathematical foundation for the spatial correlation of data among sensor nodes to form non-overlapping distributed clusters in the sensor field. Once the non-overlapping distributed clusters are constructed in the sensor field, we investigate to what degree the accurate data are extracted by sensor nodes in each distributed cluster in the sensor networks. Finally the accurate data are transmitted to the sink node by each distributed cluster in the sensor network.

2.1 Data Correlation among Sensor Nodes in Spatial Domain

We are interested to discuss a mathematical model [16] to find the spatial data correlation among sensor nodes i and j to measure a tracing point [4]. Tracing point is a reference value or an event which we are going to measure in the sensor field (spatial domain). For example, we are interested to measure the moisture content in various location of agricultural field. According to literature [3] as the sensor node density increases, the spatial correlation between observed data (S_i, S_j) among sensor nodes i and j also increases in the sensor region. The sensor nodes i as well as j can sense and measure the tracing point to collect the continuous data samples $S_i=\{ s_{i1}, s_{i2}, s_{i3}, \ldots\ldots s_{in} \}$ and $S_j=\{s_{j1}, s_{j2}, s_{j3}, \ldots\ldots s_{jn}\}$ over a window frame of time interval T.

We compute the mean, variance and covariance [16] for continuous data sample over a window frame of time interval T for the sensor nodes i and j. Finally we find the correlation coefficient $(\rho_{S_i S_j})$ between observed data (S_i, S_j) among sensor nodes i and j given by

$$\rho_{S_i S_j} = \frac{Cov(S_i S_j)}{Var(S_i)Var(S_j)} \qquad (1)$$

The spatial correlation between observed data (S_i, S_j) among sensor nodes i and j can be modeled as Joint Gaussian Random Variables (JGRV) [12, 14] as follows.

$E[S_i] = 0$, $E[S_j] = 0$; $Var[S_i] = \sigma_{si}^2$, $Var[S_j] = \sigma_{sj}^2$;

$Cov[S_i, S_j] = \sigma_{si}^2 Corr[S_i, S_j]$ for $i = 1, 2, \ldots\ldots n$ and $j = 1, 2, \ldots\ldots\ldots n$.

We define a correlation model [14] $K_V(.)$ which is given as

$$K_V(d_{i,j}) = Corr[S_i S_j] = \rho[S_i S_j] = \frac{E[S_i S_j]}{\sigma_{si}^2} = \frac{Cov[S_i S_j]}{\sigma_{si}^2} \qquad (2)$$

where $d_{i,j} = \| S_i - S_j \|$ is the Euclidian distance between the sensor nodes i and j for the sensed data. The covariance function is a non-negative function and can decrease monotonically with distance $d_{i,j} = \| S_i - S_j \|$ with limiting value of 1 at $d = 0$ and of 0 at $d = \infty$. Usually covariance models can be classified into four groups : Spherical , Power Exponential , Rational quadratic , Matern [21,22] . Out of these, power exponential model is widely used because the physical phenomenon such as electromagnetic waves can be expressed as exponential autocorrelation function [31,32]. Hence in this paper, we adopt power exponential model for correlation model $K_V(.)$ given as

$$K_V^{PE}(d) = e^{(-d/\theta_1)^{\theta_2}} \text{ for } \theta_1 > 0, \theta_2 \in (0, 2] \qquad (3)$$

We define a threshold α for $0 < \alpha \le 1$ which follows two properties:

- If $\rho_{S_i S_j} \ge \alpha_{ij}$, observed data are strongly correlated among sensor nodes i and j in the sensor field.
- If $\rho_{S_i S_j} < \alpha_{ij}$, observed data are weakly correlated among sensor nodes i and j in the sensor field.

From (1),(2) and (3) , we formulate the correlation coefficient $\rho_{S_i S_j}$ for observed data among sensor nodes i and j using power exponential model under assumption that the data are strongly correlated in the spatial domain given as follows:

$$\rho_{S_i S_j} = \frac{Cov[S_i S_j]}{\sigma_{s_i}^2} = e^{(-d_{ij}/\theta_1)^{\theta_2}} \ge \alpha \qquad (4)$$

From (4), we find the relation between the threshold value α and power exponential model given as

$$e^{(-d_{ij}/\theta_1)^{\theta_2}} \ge \alpha$$

$$d_{ij}^2 \le \theta_1^2 \sqrt[\theta_2]{\left(\log\left(\frac{1}{\alpha} \right) \right)^2} \qquad (5)$$

We compare (5) with the Euclidean distance among the coordinates of sensor nodes i and j as follows.

$$d_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2 \qquad (6)$$

From (5) and (6), we express

$$(x_i - x_j)^2 + (y_i - y_j)^2 \le \theta_1^2 \sqrt[\theta_2]{\left(\log\left(\frac{1}{\alpha} \right) \right)^2} \qquad (7)$$

Comparing (7) with the equation of a circle, we get

$$(x_i - x_j)^2 + (y_i - y_j)^2 = R^2 \qquad (8)$$

From *(7) and (8)*, we get the radius R of circular data correlation range of each sensor node i located in the sensor field represented as a centre coordinate.

$$R^2 \leq \theta_1^2 \sqrt[\theta_2]{\left(\log\left(\frac{1}{\alpha}\right)\right)^2} \qquad (9)$$

The neighboring sensor nodes j which falls under the circular data correlation area range of each sensor nodes i, the observed data among sensor nodes i and j are highly correlated in the sensor field. Radius R of data correlation range is implemented to form distributed clusters in the sensor field which is discussed bellow.

2.2 Distributed Clustering Algorithm

We form distributed non-overlapping clusters of irregular shape and size among the sensor nodes in the sensor field using a distributed clustering algorithm [16]. We construct the distributed clustering algorithm as follows:

Notations used in the algorithm:

M =Set of sensor nodes deployed in the field

$\mathbb{C}$ =Set of cluster where each element $c \in \mathbb{C}$ is of the form $c = (a_c, b_c)$ where a denotes the *CH* node of the cluster c and b denotes the associated nodes (non-CH nodes) of the cluster c.

R = Radius of data correlation range.

Distributed Clustering Algorithm
1. Start
2. Set $W = M$
3. $\forall i \in W$,let $G(i) = \{j \in W : d(i,j) \leq R, i \neq j\}$ Where $d(i,j)$ is the Euclidean distance between i and j sensor nodes.
4. $S = \{j \in M : G(j) = \max G(i)\}$, $i \in M$, we define $d_{\max}(i) = \max\limits_{j \in G(i)} d(i,j)$, where $d(i,j)$ is the Euclidian distance between nodes i and j .
5. Let $K = \arg\min\limits_{i \in S} d_{\max}(i)$
6. $\mathbb{C} = \mathbb{C} \cup \{(K, G(K)\}$
7. $W = W - \{K\} - G(K)$
8. If $W \neq \{\varnothing\}$, go to step 3
9. Stop

Finally non-overlapping distributed clusters are formed in the sensor field. Now we are interested to perform the data accuracy estimation for each distributed cluster in the sensor field which we explain in the next part.

2.3. Clustering Algorithm based Data Accuracy Estimation

As distributed non-overlapping clusters are formed in the sensor field using the clustering algorithm, we perform data accuracy [16] estimation for each distributed cluster. Each distributed cluster can measure a single tracing point of the same event in sensor field. We calculate the data accuracy for the measured data at the CH node of the respective cluster in the sensor field. The data accuracy is performed to check the estimated data received at the CH node from all the sensor nodes in the cluster are accurate and don't contain any redundant information.

We discuss the mathematical expression of data accuracy estimation for each distributed cluster with M different sensor nodes in the field. Each sensor node i can observe, sense and measure the physical phenomenon of data S_i for the tracing point value S with observation noise n_i for each distributed cluster. Therefore the observation done by the sensor node i in each distributed cluster is given as

$$x_i = s_i + n_i \quad \text{where } i \in M \qquad (10)$$

Each sensor node i sense the observed sample data x_i and transmits x_i to the CH node in each distributed cluster sharing wireless Additive White Gaussian Noise(AWGN) channel [12,23] where n_i is independent of each other and modeled as Gaussian Random Variable of zero mean and variance σ_n^2. Thus the observed sample data x_i passes through AWGN channel to the CH node for each distributed cluster in the sensor field which reconstructs estimation $\hat{S}$ of the tracing point S. The CH node receive all M observation sample for each distributed cluster given by

$$X = AZ + N \qquad (11)$$

where X is a $M \times 1$ data vector for observation done by M different sensor nodes in each distributed cluster , Z is a $(1+M) \times 1$ random vector for physically sensed data S_i for $i \in M$ including the point event S where we estimate for N $(0, C_Z)$, A is a known $M \times (1+M)$ matrix and N is a $M \times 1$ noise vector for the observed data of M different sensor nodes with N $(0, C_Z)$ in each distributed cluster in sensor filed. The random vector Z with zero mean and covariance N $(0, C_Z)$ can be shown as follows:

$$\mu_z = E[Z] = \begin{bmatrix} E[s] \\ E[s_1] \\ E[s_2] \\ . \\ . \\ E[s_M] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix} \qquad \text{and}$$

$$C_z = E[ZZ^T] = \begin{bmatrix} E[s,s] & E[ss_1] & E[ss_2] & . & . & E[ss_M] \\ E[s_1 s] & E[s_1 s_1] & E[s_1 s_2] & . & . & E[s_1 s_M] \\ E[s_2 s] & E[s_2 s_1] & E[s_2 s_2] & . & . & E[s_2 s_M] \\ . & . & . & . & . \\ . & . & . & . & . \\ E[s_M s] & . & & . & . & E[s_M s_M] \end{bmatrix}$$

Thus the covariance matrix is

$$C_z = \sigma_s^2 \begin{bmatrix} 1_{1\times 1} & r_{1\times M}^T \\ r_{M\times 1} & B_{M\times M} \end{bmatrix} \qquad (12)$$

where

$$r_{M\times 1} = \begin{bmatrix} \rho_{s_1,s} \\ \rho_{s_2,s} \\ . \\ . \\ \rho_{s_M,s} \end{bmatrix} \qquad \text{and}$$

$$B_{M\times M} = \rho_{s_i,s_j} = \begin{bmatrix} \rho_{s_1,s_1} & \rho_{s_1,s_2} & . & \rho_{s_1,s_M} \\ \rho_{s_2,s_1} & \rho_{s_2,s_2} & . & \rho_{s_2,s_M} \\ . & . & . & . \\ \rho_{s_M,s_1} & \rho_{s_M,s_2} & . & \rho_{s_M,s_M} \end{bmatrix}$$

In the matrix C_Z, $r_{M\times 1} = \rho_{S_i,S}$ illustrates the correlation coefficient between S_i , S respectively and $B_{M\times M} = \rho_{S_i,S_j}$ is the correlation coefficient between S_i , S_j respectively. The power exponential model [21,22] is used for the correlation model to show the relation between S_i and S as well as S_i and S_j . Thus we get

$$\rho_{S_i,S} = e^{-\left(d_{S,i}/\theta_1\right)^{\theta_2}} \text{ and } \quad \rho_{S_i S_j} = e^{-\left(d_{i,j}/\theta_1\right)^{\theta_2}} \text{ in } \text{ the}$$

covariance matrix C_Z .The CH node collects all the observations from M different sensor nodes in each distributed cluster to calculate the estimate of $\hat{S}$ from $\hat{S}_i$. If the observed data X can be modeled by Bayesian Linear Model [24] for all sensor nodes in each distributed cluster , the MMSE estimator to estimate the tracing point at the CH node in each cluster is given as :

$$\hat{Z} = E[Z \mid X] = \begin{bmatrix} \hat{s} \\ \hat{s}_1 \\ \hat{s}_2 \\ . \\ \hat{s}_M \end{bmatrix}$$

$$\hat{Z} = C_Z A^T (A C_Z A^T + \sigma_N^2 I_{M\times M})^{-1} X$$

$$\hat{Z} = \begin{pmatrix} r^T \\ B \end{pmatrix} \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M\times M} \right)^{-1} X \qquad (13)$$

Note : $\hat{Z} = E[Z \mid X]$ *can be represented in MMSE as well as in MAP estimator as follows :*

For MMSE : $\hat{Z} = \arg\min_Z E\left[\sum_{i=1}^{M}(Z_i - X_i)^2 \right] = E[Z \mid X]$

For MAP : $\hat{Z} = \arg\min_Z P(Z \mid X) = E[Z \mid X]$

Both MMSE and MAP have the same meaning for $E[Z \mid X]$ *in* $\hat{Z}$ *. We use MMSE estimator in this paper to find the data accuracy for each distributed cluster in the network.*

The measurement for the MMSE estimator at the CH node for each distributed cluster is given as the error $\in = (S - \hat{S})$ with mean zero and covariance matrix illustrated as

$$E[(Z - \hat{Z})(Z - \hat{Z})^T]$$
$$= C_Z - C_Z A^T (A C_Z A^T + \sigma_N^2 I_{M\times M})^{-1} A C_Z$$
$$= \sigma_S^2 \begin{pmatrix} 1 & r^T \\ r & B \end{pmatrix} - \sigma_S^2 \begin{pmatrix} r^T \\ B \end{pmatrix} \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M\times M} \right)^{-1} (r \ B) \qquad (14)$$

From (13), we calculate the estimation of tracing point $(\hat{S})$ at the CH node in each distributed cluster in the sensor region given as

$$\hat{S} = r^T \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M\times M} \right)^{-1} X \qquad (15)$$

We calculate the distortion factor between S and $\hat{S}$ to perform data accuracy estimation at the CH node for each distributed cluster in the sensor field. From equation no (14), we get the distortion factor as

$$\overline{D} = E[(S - \hat{S})^2]$$
$$\overline{D} = \sigma_S^2 - \sigma_S^2 r^T \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M\times M} \right)^{-1} r \qquad (16)$$

We normalize the distortion factor and finally calculated the data accuracy estimation of M different sensor nodes for each distributed cluster in the sensor field is given as

$$D_A(M) = 1 - \frac{\bar{D}}{\sigma_S^2}$$

$$D_A(M) = r^T \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M \times M} \right)^{-1} r \quad, \quad \text{where} \quad \left(B + \frac{\sigma_N^2}{\sigma_S^2} I_{M \times M} \right)^{-1} r = \beta$$

$$D_A(M) = r^T \beta \tag{17}$$

Data accuracy $D_A(M)$ for estimations is defined in terms of expectation of the errors between the actual point event and mean square average estimate of M sensor nodes. $D_A(M)$ calculated at the CH node for each distributed cluster in the sensor field is performed and finally transmits the most appropriate data to the sink node. Hence the purpose of verifying the data accuracy $D_A(M)$ at CH node for each distributed cluster is to confirm that the most accurate data transmitted by M different sensor node can aggregate rather than aggregating all the redundant data at the CH node [16]. The information accuracy proposed in literature [12] shows that at first each sensor nodes i can calculate the MMSE estimate $\hat{S}_i$ for observed data and then transmits the estimated data $\hat{S}_i$ to the CH node i.e. $\hat{S}_i$ in order to find $\hat{S}$. Finally averaging all $\hat{S}_i$ at the CH node for the cluster for M sensor nodes to get $\hat{S}$. But in Data accuracy estimation $D_A(M)$, at first we collect all the observed data from M sensor nodes and then only perform the MMSE estimation at the CH node for each distributed cluster. It is efficient to calculate the MMSE estimation only at the CH node rather than performing the MMSE estimation at individual nodes and then averaging it at the CH node for each distributed cluster in the sensor field.

Cluster Number	Cluster Head Node ID	Associated Nodes ID in Cluster	Information Accuracy I(M)	Data Accuracy D(M)
I	14	12,16,17,23,24,26,28,29	0.8502	0.9577
II	1	5,7,11,13,15,18,19	0.8690	0.9632
III	10	2,3,8,20,21,27,30	0.8769	0.9585
IV	9	4,25	0.8732	0.9459
V	6	22	0.8349	0.9627

Table 1: Distributed clusters with data accuracy estimation

According to the clustering algorithm discussed in Section.2.2 with threshold value $\alpha = 0.7$, we form five non

overlapping distributed clusters of different sizes after performing simulation as shown in Table.1. Each distributed cluster has its CH node with associated sensor nodes which perform the data accuracy at the CH node. We compare the degree of accuracy for information accuracy model [12] with our data accuracy $D_A(M)$ estimation which shows $D_A(M)$ always perform better than $I(M)$ for each distributed cluster in the network.

We find that each distributed cluster in the network perform data accuracy $D_A(M)$ at their respective CH node and transmits the most accurate data to the sink node. We take the first cluster having nine sensor nodes given in Table.1 which we get after simulating the clustering algorithm. Fig.1, shows that six sensor nodes are sufficient to perform the same data accuracy level achieve by nine sensor nodes in the cluster with $\theta_1 = 70$ and $\theta_2 = 1$. Hence, we only choose certain number of sensor nodes which are close to the tracing point instead of choosing all sensor nodes in each distributed cluster subjected to achieve approximately good data accuracy in the network. Thus we can reduce the number of sensor nodes in each distributed cluster subjected to data accuracy estimation in the sensor network.
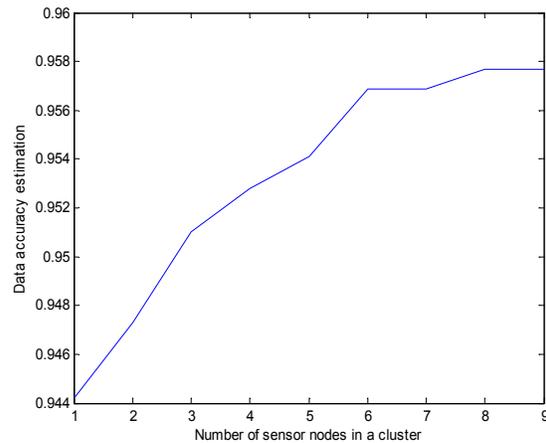

Figure 1: Number of sensor nodes vs. data accuracy in a cluster

3. Probabilistic Model for Node Selection in the Network

In the previous section, we find the data accuracy for each distributed cluster and conclude that there exists an optimal number of sensor nodes in each distributed cluster. Moreover these optimal sensor nodes in each distributed cluster are sufficient to perform approximately the same data accuracy achieve by each cluster. Hence it is unnecessary to choose all the sensor nodes in each distributed cluster to perform the transmission process in

the network. Optimal number of sensor nodes which are active can perform the data accuracy and rest of the sensor nodes goes to sleep mode in each distributed cluster in the network. Since an optimal number of sensor nodes are active and rests of them are in sleep mode in each distributed cluster, we perform an analytical as well as simulation model for energy consumption in the networks to save energy. Each distributed cluster is independent to each other irrespective of transmission of data packets to the sink node. Each distributed cluster has its own intelligence to operate its associated sensor nodes to switch on (active mode) or switch off (sleep mode). Since each distributed cluster has its own control to operate the associate sensor nodes, we find the probability by which the sensor nodes are in active mode and sleep mode subjected to data accuracy and energy consumption.

3.1. Probability by which Sensor Nodes are Active Subjected to Data Accuracy

Here we perform analytical and simulation model for the sensor nodes which are active subjected to data accuracy in each distributed cluster in the network. We demonstrate a probabilistic model for sensor nodes which are in active mode and rest of the sensor nodes are in sleep mode in each distributed cluster in the network. Since in each distributed cluster in the network, CH node performs the data accuracy for associated nodes in the cluster, CH node is always switched on. The associated nodes are in active mode with probability P and in sleep mode with probability $(1-P)$ in each distributed cluster in the network. We go through the following steps to validate our model:

M= Total number of sensor nodes in each distributed cluster in the network
N= Total number of sensor nodes which are switch on (active mode) in each distributed cluster in the network
X_0 =CH node always switches on in each distributed cluster
$m*$ =Expected number of sensor nodes which are in active mode in each distributed network

Total numbers of sensor nodes which are in active mode for each distributed cluster in the network are given as

$$N = X_0 + \overline{X}$$
$$N = X_0 + \sum_{i=1}^{M-1} X_i \qquad (18)$$

Where
$X_i = 1$, if nodes i selected with probability P, (sensor nodes with switch on)
$X_i = 0$, if nodes i selected with probability $(1-P)$, (sensor nodes with switch off)

Expected number of sensor nodes which are in active mode (switch on) with probability P in each distributed cluster in the network is given as

$$E(N) = E(X_0) + \sum_{i=1}^{M-1} E(X_i)$$
$$E(N) = 1 + (M-1)P$$
$$E(N) = MP - P + 1$$
$$E[N] = m* \qquad (19)$$

Probability by which the sensor nodes are in active mode (switch on) in each distributed cluster in the network is given as

$$P = \frac{m*-1}{M-1} \qquad (20)$$

We find all the combinations of M sensor nodes taken K at a time in each distributed cluster which are given as $^{M}C_K$. Since the CH node is fixed for every combination in each distributed cluster in the network, the combination becomes $^{M-1}C_{K-1}$ for the number of all the combinations of M sensor nodes taken K at a time given as.

$$^{M}C_K = \frac{M!}{K!(M-K)!}$$

$$= \frac{M(M-1)!}{K(K-1)!((M-1)-(k-1))!}$$

$$= \frac{M}{K}{}^{M-1}C_{K-1}$$

$$^{M-1}C_{K-1} = \frac{K}{M}C_K^M \qquad (21)$$

Proof that: $\quad ^{M-1}C_{K-1} = \frac{K}{M}C_K^M$

$$^{M-1}C_{K-1} = \frac{M(M-1)!}{M(K-1)!((M-1)-(K-1))!}$$

$$= \frac{M!}{M(K-1)!(M-K)!}$$

$$= \frac{(M-K+1)M!}{M(M-K+1)(K-1)!(M-K)!}$$

$$= \frac{(M-K+1)}{M}\left[\frac{M!}{(K-1)!(M-K+1)!}\right]$$

$$= \frac{(M-K+1)}{M}\left[\frac{M!}{(K-1)!(M-K+1)(M-K)!}\right]$$

$$= \frac{K}{M}\left[\frac{M!}{K(K-1)!(M-K)!}\right]$$

$$= \frac{K}{M}\left[\frac{M!}{K!(M-K)!}\right]$$

$$= \frac{K}{M}{}^{M}C_K$$

The probability mass function (pmf) for active nodes in each distributed cluster with a fixed CH node in the network can be represented by binomial distribution given as.

$$P(N = K) = P(\overline{X} = K - 1)$$
$$= {}^{M-1}C_{K-1}P^{K-1}(1-P)^{M-K}$$
$$= \frac{K}{M}{}^{M}C_{K}P^{K-1}(1-P)^{M-K} \qquad (22)$$

The probability of sensor nodes which are in active mode subjected to data accuracy function $A(P)$ in each distributed cluster in the network is given by

$$A(P) = \sum_{K=1}^{M} \frac{K}{M} A(N = K) {}^{M}C_{K}P^{K-1}(1-P)^{M-K} \qquad (23)$$

Now we find a minimum probability ($P_{\min}$) with χ_a as a user dependent factor $(0 < \chi_a < 1)$ to switch on the sensor nodes in the network. Thus the minimum probability for achieving a required level of data accuracy to switch on the sensor nodes in each distributed clusters is given as

$$P_{\min} = \arg\min_{P}\{A(P) \geq \chi_a A_{\max}\} \quad \text{where} \quad A_{\max} = A(1) \quad (24)$$
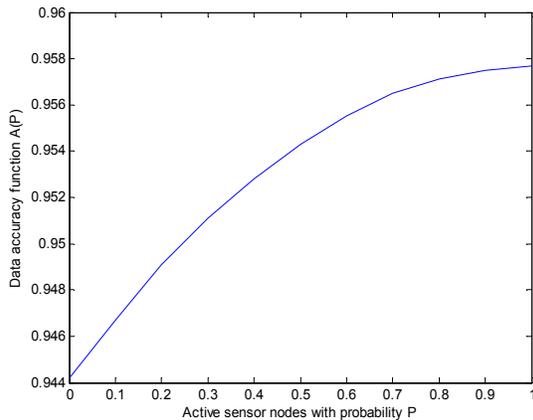
Figure 2: Data accuracy function vs active sensor nodes with probability P

By implementing the clustering algorithm with threshold value α =0.7 discussed in Section.2.2, we form distributed non overlapping clusters with different sizes in the network. We choose the first cluster with nine sensor nodes from the network to satisfy (23), which performs the data accuracy for this cluster as shown in Fig. 2. We illustrate the analysis of this cluster in probabilistic form subjected to data accuracy function. In Fig. 2, we plot the probability of sensor nodes which are in active mode for the cluster with respect to data accuracy. We take minimum probability $P_{\min}$ =0.7 for achieving a required level of data accuracy $A(P)$ for χ_a =0.956 to switch on the sensor nodes in the cluster. So with $P_{\min}$ a certain

number of sensor nodes are switched on which are sufficient to give approximately the same data accuracy level achieve by the all the sensor nodes in each distributed cluster. Thus each an every distributed cluster can switch on certain number of sensor nodes (generally sensor nodes close to tracing point) with minimum probability $P_{\min}$ and rest of them goes to sleep mode in each distributed cluster in the network. Thus an optimal number of sensor nodes in each distributed cluster are sufficient to operate with $P_{\min}$ in the networks respect to data accuracy instead of switch on all the sensor nodes in the networks. Hence we can save node deployment cost per cluster by operating $P_{\min}$ in different distributed cluster in the network.

3.2. Probability by which Sensor Nodes are Active Subjected to Energy Consumption

Each distributed cluster can perform the energy consumption model [6] in the network. We analyze and simulate a probabilistic model for energy consumption of each distributed cluster in the network. For our simplicity we model the energy consumption for each distributed cluster and then implement it to calculate the energy consumption for the whole network. For each distributed cluster in the network, energy consumption is given as

$$E_{Cluster} = E_{CH} + E_{Non-CH}(M - 1) \qquad (25)$$

where E_{CH} is the energy consumed for the CH node, the energy consumed for non-CH nodes is given by E_{Non-CH} and M is the total number of sensor nodes in each distributed cluster.

We take a radio hardware energy dissipation model [6], in which transmitter antenna dissipates energy for the radio electronics, power amplifier and the receiver antenna to run the radio electronics. We use both the free space (d^2 power loss) channel model and the multipath fading (d^4 power loss) channel model which is dependent on the distance between transmitter and receiver antenna respectively [30].

We define a threshold τ_0, if the distance is less than the threshold τ_0, the free space model is used otherwise we can use multipath model [6]. Hence the energy consumption to transmit the data is given as

$$E_{TX} = lE_{Elec} + l \in_{fs} d^2 \qquad \text{for } d < \tau_0$$
$$= lE_{Elec} + l \in_{mp} d^4 \qquad \text{for } d \geq \tau_0 \quad (26)$$

Energy consumed to receive the data is given as $E_{RX} = lE_{Elec}$. E_{Elec} is a electronics energy which depends upon spreading of signal, modulation techniques and coding to be used. $\in_{fs} d^2$ and $\in_{mp} d^4$ are the amplifier energy depends upon the distance between the transmitter and the receiver antenna respectively. The parameters for

communication energy are given as: $E_{Elec} = 50nJ/bit$, $\in_{fs} = 10pJ/bit/m^2$, $\in_{mp} = 0.0013pJ /bit/ m^4$ and the data aggregation is set as $E_{Agg} = 5nJ/bit /signal$.

Each non-CH node in a cluster only transmits the data to the CH node. Moreover the distance between the non-CH nodes to the CH node is small in each distributed cluster in the network. Hence we adopt free space channel model (d^2 power loss).Thus the energy consumption in each non-CH node in a cluster is given as

$$E_{Non-CH} = E_{Tx}$$
$$= lE_{Elec} + l \in_{fs} d_{toCH}^2 \qquad (27)$$

We define the sensor nodes (non-CH nodes) at a polar coordinate (r, θ) from the CH node in a cluster and integrate over the disk as follows.

$$E_{non-CH} = lE_{Elec} + l \in_{fs} \int_0^{2\pi} \int_0^R r^2 \rho \, dr d\theta$$

Where R is the radius of the correlation coefficient of a node discussed in Section. 2.1 and $\rho = \dfrac{M}{\pi R^2}$.

$$E_{non-CH} = lE_{Elec} + l \in_{fs} \left(\frac{MR^2}{2} \right) \qquad (28)$$

M and R varies in each distributed cluster in the network depending upon the cluster size. The energy consumption at the CH node for each distributed cluster can be given as receiving the data from the non-CH nodes, aggregating the data and transmitting the aggregate data to the base station. Since the base station is far away from each distributed cluster, we use the multipath channel model (d^4 power loss). Thus the energy consumed at the CH node in each distributed cluster is given as:

$$E_{CH} = E_{RX} + E_{Agg} + E_{TX}$$
$$= lE_{Elec}(M-1) + lE_{Agg}M + lE_{Elec} + l \in_{mp} d_{toBS}^4 \qquad (29)$$

From (28) and (29), the total energy consumption in each distributed cluster in the network is given as

$$E_{cluster} = E_{CH} + (M-1)E_{non-CH} \qquad (30)$$

Now we develop a probabilistic model for active sensor nodes in each distributed cluster subjected to energy consumption. The probability by which energy is consumed in each distributed cluster in the network is given as

$$E_{Cluster}(P) = \sum_{K=1}^{M} E_{Cluster}(N=K)P(N=K)$$

$$= \sum_{K=1}^{M} E_{Cluster}(N=K) \left[{}^{M-1}C_{K-1} P^{K-1}(1-P)^{M-K} \right]$$
$$= \sum_{K=1}^{M} E_{Cluster} \left[\left(\frac{K}{M} C_K^M \right) P^{K-1}(1-P)^{M-K} \right] \qquad (31)$$

Now we define a maximum probability P_{max} with χ_e a user dependent factor ($0 < \chi_e < 1$) to switch on the sensor nodes in each distributed cluster subjected to energy consumption. Thus the maximum probability P_{max} for keeping the energy consumption below a certain level to switch on the sensor nodes in each distributed cluster in the network is given as

$$P_{max} = \arg\max_P \{E_{cluster}(P) \le \chi_e E_{max}\} \text{ where } E_{max} = E_{cluster}(1) \quad (32)$$
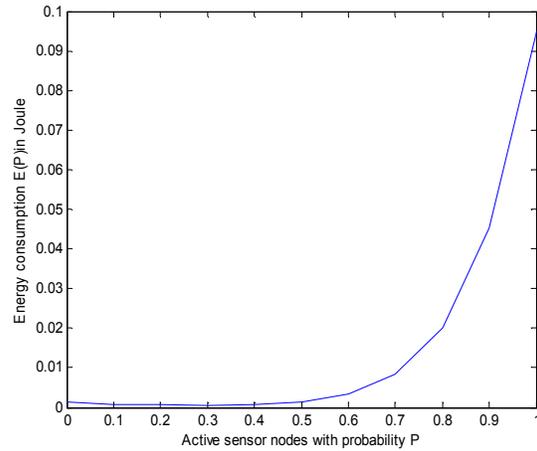

Figure 3: Energy consumption vs. active sensor nodes with probability P

We take the first cluster with nine sensor nodes chosen from the clustering algorithm with threshold value $\alpha = 0.7$ and perform the energy consumption model for the cluster. We analysis the energy consumption for the cluster in probabilistic form. In Fig. 3, we plot the probability of the sensor nodes which are switched on for the cluster of nine sensor nodes subjected to energy consumption. We take maximum probability $P_{max} = 0.6$ to keep the energy consumption $E_{cluster}(P)$ for $\chi_e = 0.004$ below a certain level to switch on the sensor nodes in the cluster. Thus it is clear from the Fig. 3 that we can operate P_{max} such that certain numbers of sensor nodes are switch on subjected to energy consumption rather than switching all the sensor nodes in each distributed cluster. Thus we save the energy cost by reducing the number of sensor nodes in each distributed clusters and finally optimized the energy consumption in the whole network by operating P_{max}.

3.3. Trade Off Between Data Accuracy and Energy Consumption for Node Selection in the Network

Number of clusters	Total No. of nodes in each cluster	Probability P for selecting nodes	Expected No. of nodes E[N]=m* selected in each cluster
I	9	0.65	6
II	8	0.5	5
III	8	0.6	5
IV	3	0.4	2
V	2	0.5	2
Total no. of nodes in network	30 Nodes		20 Nodes Selected in network

Table 2: Expected number of sensor nodes selected in the network

Previously we have illustrated that the sensor nodes are switched on and switched off in each distributed cluster in the network subjected to data accuracy and energy consumption in probabilistic approach. Thus it leads to the trade off between P_{min} for data accuracy and P_{max} for energy consumption for selecting an optimal sensor nodes in each distributed cluster in the network. Hence we define a probability P such that $P \in [(P_{min} + P_{max})/2]$ to select an optimal number of sensor nodes which are sufficient to achieve the data accuracy $A(P)$ and can reduce the energy consumption $E_{cluster}(P)$ in each distributed cluster in the sensor network. The probability P for the cluster of nine sensor nodes discussed in Section.3.1 and Section.3.2 is 0.65 and the expected numbers of sensor nodes selected in the first cluster are six out of nine sensor nodes. Hence we find an optimal expected number of sensor nodes by fixing an appropriate probabilistic value P for active nodes in each distributed cluster having trade off between data accuracy and energy consumption. This leads to reduce the node deployment cost and energy cost per cluster in the distributed network.
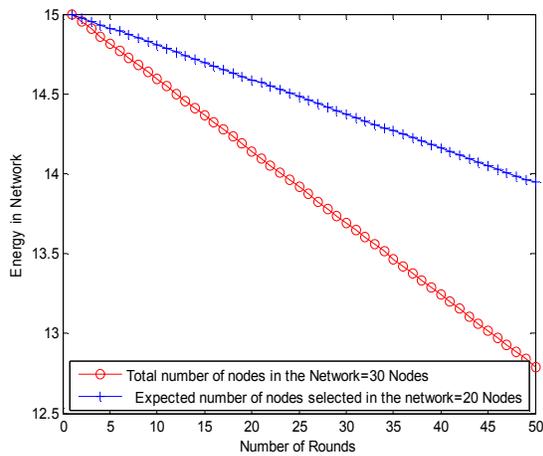

Figure 4: Energy in network vs. number of rounds

In Table. 2, we are interested to show the expected number of sensor nodes selected in the network. Initially we deploy thirty sensor nodes in the sensor field. According to the clustering algorithm discussed in Section.2.2, sensor nodes form five distributed non-overlapping clusters in the sensor field. Each distributed cluster fixes a different probabilistic value P to select active sensor nodes depending upon the trade off between data accuracy and energy consumption. Hence out of thirty sensor nodes, twenty sensor nodes are selected for data communication in the network. Finally we simulate energy consumption model for thirty sensor nodes and selected twenty sensor nodes as shown in Fig. 4. We conclude that energy consumption for thirty sensor nodes decays much faster than the selected twenty sensor nodes in the network with respect to time. Thus reducing the number of sensor nodes can save energy consumption in the whole network and increases life time of the network.

4. Conclusions

We conclude that we find the optimal number of sensor nodes using the trade off between data accuracy and energy consumption in probabilistic approach in each distributed cluster. This reduces the node deployment cost and energy cost in the network. We illustrate the probabilistic approach by which the sensor nodes can be operated for active mode and sleep mode using the trade off between data accuracy and energy consumption in each distributed cluster in the network. Simulation results shows energy consumption for the total number of sensor nodes deployed in the sensor region decays much faster than the optimal number of sensor nodes selected with respect to time. This increases life time of the network. Moreover our data accuracy estimation model performs better than information accuracy model in each distributed cluster with respect to data accuracy.
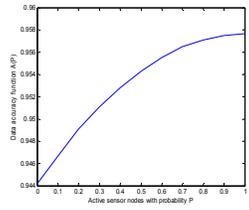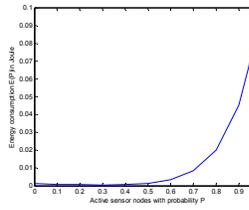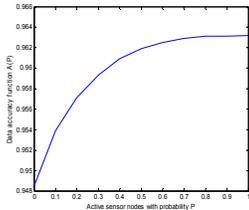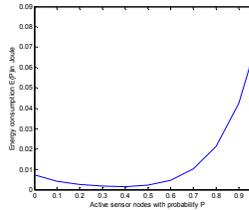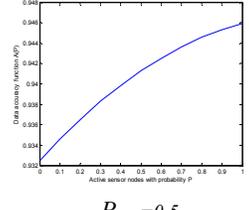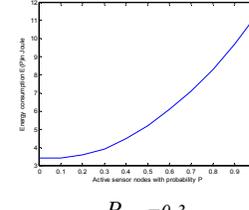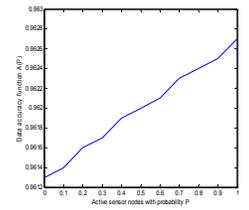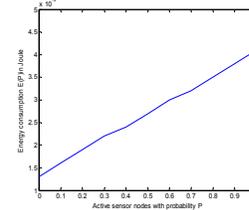
References

[1] I.F Akyuildz ,W.Su, Y.Sankarasubramanian and E. Cayirci,"A Survey on Sensor Networks ", IEEE Communications Magazine ,vol.40,pp.102-114.Aug 2002.

[2] S.S. Pradhan ,K.Ramchandran ,"Distributed Source Coding : Symmetric Rates and Applications to Sensor Networks", in the proceedings of the data compressions conference ,pp.363-372, 2002.

[3] C.Zhang ,B.Wang ,S.Fang,Z Li ,"Clustering Algorithm for Wireless Sensor Networks using Spatial Data Correlation ", Proceedings of IEEE International Conference on Information and Automation ,pp.53-58,June 2008.

[4] Jyotirmoy Karjee , H.S Jamadagni ," Data Accuracy Estimation for Spatially Correlated Data in Wireless Sensor Networks under Distributed Clustering ", Journal of Networks,vol-6,no.7,pp1072-1083,July 2011.

[5] A.Abbasi and M.Younis ,"A Survey on Clustering Algorithms for Wireless Sensor Networks ", Computer Communications ,vol-30,no.14-15,pp2826-2841,2007.

[6] W.B Heinzelman , Anantha P. Chandrakasan , " An Application Specific Protocol Architecture for Wireless Microsensor Networks", IEEE transactions on Wireless Communications , vol-1 , no. 4 , pp. 660-670,Oct 2002.

[7] L. Guo , F chen, Z Dai , Z. Liu," Wireless Sensor Network Cluster Head Selection Algorithm based on Neural Networks ", International Conference on Machine vision and human machine Interference , pp.258-260,2010.

[8] Georgios Smaragdakis ,Ibrahim Matta ,Azer Bestavros, "SEP: A stable Election Protocol for Cluster Heterogeneous Wireless Sensor Networks".

[9] Chongqing Zhang ,Binguo Wang , Sheng Fang , Jiye Zheng, " Spatial Data Correlation Based Clustering Algorithms for Wireless Sensor Networks", The 3rd International Conference on Innovative Computing Information and Control (ICICIC'08).

[10] Zhikui chen , Song Yang , Liang Li and Zhijiang Xie ," A clustering Approximation Mechanism based on Data Spatial Correlation in Wireless Sensor Networks", Proceedings of the 9th International Conference on Wireless Telecommunication Symposium 2010.

[11] Ali Dabirmoghaddam ,Majid Ghaderi ,Carey Williamson, " Energy Efficient Clustering in Wireless sensor Networks with Spatially Correlated Data", IEEE Infocom 2010 proceedings.

[12] Kang Cai, Gang Wei and Huifang Li, "Information Accuracy versus Jointly Sensing Nodes in Wireless Sensor Networks ",IEEE Asia Pacific Conference on Circuit and System ,pp.1050-1053,2008.

[13] M.Gastpar ,M.Vetterli , "Source Channel Communications in Sensor Networks", Second International Workshop on Information Processing in Sensor Networks ,(IPSN'2003).

[14] Varun M.C,Akan O.B and I.F Akyildz, "Spatio – Temporal Correlation : Theory and Applications in Wireless Sensor Networks", Computer Network Journal (Elsevier Science), vol. 45,pp.245-259,june 2004.

[15] Jyotirmoy Karjee,H.S Jamadagni, " Data Accuracy Estimation for Cluster with Spatially Correlated Data in Wireless Sensor Networks", IEEE International Conference on Information System and Computational Intelligence , vol-3,pp.284-291,Harbin , China, 2011.

[16] Jyotirmoy karjee, H.S Jamadagni ," Data Accuracy Model for Distributed Clustering Algorithm based on Spatial Data Correlation in Wireless Sensor Networks" uploaded in arxiv.org (http://arxiv.org/abs/1108.2644)

[17] Zhang Hu, Zhang Huiyan ,"Node Selection Algoritm Optimized for wireless Sensor networks",workshop on knowledge and data mining , 2008.

[18] Jyotirmoy karjee, Sudipto Banerjee, "Tracing the Abnormal Behavior of Malicious Nodes in MANET ", Fourth International Conference on Wireless Communications, Networking and Mobile Computing, pp.1-7, Dalian , China ,2008.

[19] T.Minming ,N Jieru, W Hu, Liu Xiaowen ,"A data Aggregation Model for underground Wireless Sensor Network", WRI World Congress on Computer Science and Information Engineering, vol -1,pp.344-348,2009.

[20] Huifang Li, Shengming Jiang , Gang Wei, "Information Accuracy Aware Jointly Sensing Nodes Selection in Wireless Sensor Networks", LNCS 4325.pp. 736-747, MSN Springer -2006.

[21] J.O Berger, V.de Oliviera and B.Sanso , " Objective Bayesian Analysis of Spatially Correlated Data ", Journal of Am. Statistics. Assoc., vol-96, pp.1361-1374, 2001.

[22] De Oliveria V, Kedan B and Short D.A ,"Bayesian Predication of Transformed Gaussian Random Fields ", Journal of American Statistical Association , pp. 1422-1433,1992.

[23] T.J Goblick, "Theoretical Limitations on the transmission of data from Analog Sources". IEEE Transaction Theory, IT-11(4), pp.558-567,1965.

[24] Steven M. Kay, "Fundamentals of Statistical Signal Processing –Estimation Theory", Pearson, Volume-1,2010.

[25] V.Poor ,"An Introduction to Signal Detection and Estimation ",Second Edition , Springer , Berlin 1994.

[26] C.Y. Cho, C.L Lin , Y.H Hsiao, J S Wang , K.C Yong , " Data Aggregation with Spatially Correlated Grouping Techniques on Cluster Based WSNs", SENSOR COMM, pp-584-589, Venice -2010.

[27] Shirshu Varma , Uma Shankar Tiwary, " Data Aggregation in Cluster Based Wireless Sensor Networks", Proceedings of the first International Conference on Intelligent Human Computer Interaction , pp.391-400, part-5,2009.

[28] S.Soro,Wedi B. Heinzelman ,"Cluster Head Election Techniques for coverage Preservation in Wireless Sensor Networks", Adhoc Networks, Elsevier, pp-955-972,2009.

[29] D.Tian , N.Georganas ,"A Node Scheduling Scheme foe Energy Conservation in large Wireless Sensor Networks ",Wireless Communications and Mobile Computing Journal,3(2):271-290, March 2003.

[30] T. Rappaport, Wireless Communications : Principles & Practice . Englewood Cliffs , NJ : Prentice Hall 1996.

[31] M.C Vuran , I.F Akyildiz , "Spatial Correlation – based Collaborative Medium Access Control in Wireless Sensor Networks", IEEE/ACM Transactions on Networking,2005.

[32] G.L Stuber, " Principles of mobile communication ", Kluwer Academic publisher ,2001.

Jyotirmoy Karjee received his B.E (Electronics), M.E(Information Technology) specialization in Network Security in 2003 and 2005 respectively. He worked in Prakriti Inbound Pvt. Ltd as a software engineer for a year and worked as a lecturer in Sikkim Manipal Institute of Technology, Sikkim till 2008. He is currently pursuing his Ph.D degree at Centre for Electronics Design and Technology, Indian Institute of Science, Bangalore. His current research interest includes data accuracy estimation and data aggregation in wireless sensor networks.

Prof. H.S Jamadagni received his M.E and Ph.D degree in Electrical & Communication Engineering from Indian Institute of Science, Bangalore. Currently He is the professor at Centre for Electronics Design and Technology, Indian Institute of Science. He is one of the main coordinators for the intel higher education program and was the key mentors for various intel workshops in india. His current research work includes in the areas of embedded systems, VLSI for wireless networks and wireless sensor networks.

Appendix-I: Summery for selecting optimal sensor nodes in each distributed cluster using trade off between data accuracy and energy consumption.
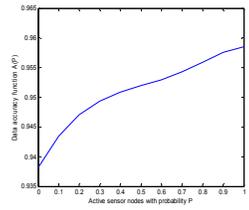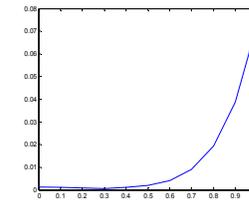
Number of clusters	No. of nodes in each cluster	P_{min} for Data accuracy	P_{max} for Energy consumption	Probability P for selecting nodes	Expected No. of nodes E[N]=m* selected
I	9	$P_{min}=0.7$	$P_{max}=0.6$	0.65	6
II	8	$P_{min}=0.5$	$P_{max}=0.5$	0.5	5
III	8	$P_{min}=0.8$	$P_{max}=0.4$	0.6	5
IV	3	$P_{min}=0.5$	$P_{max}=0.3$	0.4	2
V	2	$P_{min}=0.5$	$P_{max}=0.5$	0.5	2
Total nodes in Network	30 Nodes				20 Nodes selected