ECODA - Entropy Correlation Based Data Aggregation Protocol For Wireless Sensor Networks

Nga Nguyen Thi Thanh* and Khanh Nguyen Kim

Abstract— *Correlation characteristics can bring many significant potential advantages for the development of efficient communication protocols for wireless sensor networks. To exploit the correlation in WSNs, it is necessary to build the correlation model and developed the correlation-based protocol using this correlation model. However, most of the present correlation-based protocols only consider the linear and distance dependence correlation or computation complexity. This paper proposes the Entropy Correlation Based Data Aggregation (ECODA) Protocol for wireless sensor networks with less computation complexity that could be applied practically. This protocol uses a general correlation model with less computation complexity. In addition, two energy-efficient aggregation schemes including an on-off scheme which offers an efficient way to choose representative nodes in a cluster with permitted distortion and compression scheme which reduces in-network message length suitable to high correlation data are used in this protocol. Simulations show the effectiveness of the proposed protocol.*

Keywords— **Entropy correlation coefficient, correlation model, compression, representative node, distortion.**

1. INTRODUCTION

Because of low-cost, small in size of sensor nodes, wireless sensor networks (WSN), which expand sensing capabilities in space and time are widely used in various modern applications. However, since most of the sensor nodes are powered by no-replace batteries, energy conservation is commonly recognized as the key challenge in designing and operating the network.

In typical WSNs applications, sensors are required spatially dense deployment in order to achieve satisfactory coverage [1]. As a consequence, their sensed data are correlated with each other. The existence of correlation characteristics can bring many significant potential advantages for the development of efficient communication protocols well-suited for the WSNs paradigm [2, 3].

To exploit the correlation in WSNs, there have been many research efforts to study the correlation model and develop correlation-based protocols in WSNs. In [3], correlated nodes are supposed to observe the same source. Thus, the correlation relation is distance's dependence and could be classified into four groups including Spherical, Power exponential, Rational quadratic and Matérn. In [4], the correlation coefficient is also a function of distance among nodes. Other researches consider the correlation as the similarity of sensed data [5]. Some papers define the correlation relation in different ways such as a linear predictive model [6], node weight [7], data density correlation degree [8].

All the above researches consider only the linear correlation between data and distance-based. To solve a more general correlation relation, entropy-based correlation is considered [9- 12]. In [9], the joint entropy of a group of nodes is calculated using real data set and then a distance-based joint entropy function is built by approximation to the calculated joint entropy. Distance-based joint entropy models are proposed in [9, 10]. In [11], instead of calculating directly from real data, the entropy correlation coefficient is chosen to be the Pearson linear correlation coefficient to reduce the computation complexity but reduce the generality of using entropy. In [12], joint entropy is calculated from real data and then the joint entropy of a node-set is approximated by an exponential function of a number of nodes in the set. The advantage of this model is a distance-independent model, but the disadvantage is the complexity in the determination correlation among nodes. Joint entropy values of all possible node groups have to be calculated in order to select correlated nodes.

To overcome these above difficulties, the authors have proposed a novel correlation model using the concept of correlation ratio in [13, 14]. This model has been used to evaluate the impact of correlation to data aggregation in WSNs. In this paper, based on the result in [13, 14], we propose a novel routing protocol called ECODA (**E**ntropy **CO**rrelation Based **D**ata **A**ggregation Protocol) for WSNs. In this protocol, the clustering is based on the correlation between collected data from sensors deployed in the observed field. After the clustering process, the sensor nodes are divided into various clusters in which their data is correlated with each other by an entropy correlation coefficient. In each cluster, because of the high correlation of the data, there is redundant information if every node sends data to cluster head regularly which causes a shortage in network system lifetime. By deploying the aggregation methods which are presented in [13, 14], the network system lifetime is prolonged while satisfying the demand distortion.

The rest of the paper is presented as follows. In section

*Nga Nguyen Thi Thanh and Khanh Nguyen Kim are with Hanoi University of Science and Technology, 01 Dai Co Viet Road, Hai Ba Trung District, Ha Noi, Viet Nam.

*Corresponding author: Nga Nguyen Thi Thanh, Phone +84-24-38696125; E-mail: ngantt@soict.hust.edu.vn.

2, the correlation model is presented. Data aggregation is presented in section 3. The outline of the ECODA protocol is then shown in section 4. In section 5, the performance of the ECODA is shown. Conclusion and further study are shown in section 6.

2. ENTROPY CORRELATION

In this section, we review the results of the correlation that has been proposed in [13, 14]. These results will be used to build the proposed protocol.

Correlation region definition

Definition 1: A group of m nodes $\{X_1, X_2, \dots, X_m\}$ is in a correlation region with correlation level ρ_0 and this correlation region is the one in which the sensed data of all sensor nodes have the same entropy value. In addition, the entropy correlation coefficient between all pairs of nodes are also the same and equal to ρ_0.

$$H_0 = H(X_1) = H(X_2) = \cdots = H(X_m) \quad (1)$$

$$\rho_0 = \rho_{ij} = \rho(X_i, X_j), \forall\, i \neq j. \quad (2)$$

However, in practical cases, it is difficult to obtain the same entropy value of nodes or the same entropy correlation coefficient of pairs of nodes. Then, the correlation region can be defined in a more practical way as bellows.

Definition 2: A group of m nodes $\{X_1, X_2, \dots, X_m\}$ is in a correlation region with correlation level ρ_0 if entropies of all member nodes vary in a very small range and entropy correlation coefficients between all pairs of nodes are larger than or equal to ρ_0.

$$H_0 - \Delta H \leq H(X_1), H(X_2), \dots, H(X_m) \leq H_0, \quad (3)$$

$$\rho_0 \leq \rho_{ij} = \rho(X_i, X_j), \forall\, i \neq j \quad (4)$$

in which ΔH is the entropy variation range, H_0 is called "base entropy" and ρ_0 is called the "correlation level" of the data collected in the region. The higher the correlation level is, the more the correlation of the collected data in this region is. In this paper, if a region has $\rho_0 \geq 0.5$, we call it is a highly correlated region.

Correlation clustering algorithm

Using the definition of correlation region, a sensor field can be divided into correlation regions with specified base entropy and correlation level. The clustering algorithm is described in Fig. 1. At first, an entropy range and correlation level is chosen. Next, nodes with their entropy values in entropy range are selected into a group. Then, the entropy correlation coefficients of all pairs in the group are calculated and a node with the highest number of pairs that satisfied the correlation level is chosen as a core node. Nodes in the group that their correlation coefficients with the core node are smaller than the correlation level will be removed from the group first. After that, the process of removing a node with the highest number of pairs that do not satisfy the correlation level is repeated until all pair in the group satisfies the correlation level.

```
1  BEGIN
2    REPEAT
3    Choose H0, ρ0, ΔH; (*)
4    Initialize new group G = ∅;
5    FOR each node Xi not belong to any
group
6        IF H0 − ΔH ≤ H(Xi) ≤ H0
7           Add Xi into G
8        END_IF
9     END_FOR
10    FOR each node Xi in G
11      B(Xi)= number of nodes Xj that
        ρij ≥ ρ0
12    END_FOR
13    X0 = argmax{B(Xi), Xi∈G}
14    FOR each node Xi in G
15       IF ρ(Xi,X0) < ρ0
16          Remove Xi from G
17       END_IF
18    END_FOR
19    REPEAT
20       FOR each node Xi in G
21        C(Xi)= number of nodes Xj
          that ρij < ρ0
22       END_FOR
23       FOR each node Xi in G
24          IF  0 < C(Xi) = max{C(Xj), Xj∈G}
25             Remove Xi from G (**)
26          END_IF
27       END_FOR
28    UNTIL max{C(Xj), Xj∈G} =0
29 UNTIL all nodes are grouped
30 END
```

Fig. 1. Correlation-based clustering algorithm.

In this algorithm, G presented for a correlated group, and it is equivalent to a cluster of the network. This algorithm is implemented at the base station, and the base station records this group and then does the network clustering.

3. DATA AGGREGATION

Entropy correlation can allow efficient data aggregation. In this paper, we consider two types of data aggregations including data compression and representative types. According to [13] about data compression and [14] about representative aggregation, we conclude the results as follows.

Compression aggregation

In correlation networks, nodes are divided into correlated regions. In order to reduce the amount of data transmission, data compression can be done at some nodes in correlation regions during data transmission to the base station. And the optimal routing scheme in correlation networks can be established as follows:

- If compression along the transmission path to the cluster head is used, it is not necessary to divide a correlated region into smaller clusters to optimize the transmission cost. Instead, each correlated region becomes a cluster, and the optimal routing path in each cluster is the shortest path from nodes to their cluster head.
- If compression is done at the cluster head only, not at intermediate nodes, the transmission path is the shortest path to the cluster head. To get optimal transmission costs, it is necessary to divide a correlated region into some smaller clusters. It is difficult to get the analytical solution of optimal cluster size. But we can draw the total transmission cost curves and find out the near-optimal value with a specified correlation coefficient and the number of network nodes.

Representative aggregation

In a correlation region with a high enough correlation level, it may not be necessary for every sensor node in a correlation group to transmit its data to the base station. Instead, a smaller number of sensor measurements might be adequate to communicate the event features to the base station within a certain reliability/fidelity level. These working sensors are called representative nodes of the region/group. To evaluate the reliability/fidelity level, the distortion function is used.

In order to use representative aggregation, at first, the number of representative nodes of each correlation cluster is determined. This number depends on the entropy correlation coefficient of the correlation region and the desired distortion. The calculation of the number of representative nodes can be seen in [14].

After knowing the number of representative nodes, it is necessary to select these nodes in the cluster group. The selection can be based on different purposes such as maximizing the total information (the obtained information from representative nodes is maximum), maximizing coverage (total covered areas by representative nodes is maximum) or energy balancing (the nodes with highest remaining energy are chosen to be representative nodes).

4. ENTROPY CORRELATION BASED DATA AGGREGATION PROTOCOL (ECODA)

Because ECODA uses the entropy correlation-based clustering scheme, it is necessary to calculate the entropy of each node to estimate the joint entropy for the clustering process. However, calculating entropy requires a certain processing ability in which a single sensor node could not have response-ability. On the other hand, from the beginning, there are no data collected from sensor nodes providing for the entropy calculation. Therefore, ECODA has some following characteristics:

- The clustering process is performed by the base station because of its high processing capacity with an unlimited energy resource. On the other hand, the base station knows the information of all nodes in the network which a single node does not have.
- At least N_s samples are necessary to calculate entropies and correlation coefficients. Therefore, the operation of the protocol needs a period to collect data serving for correlation calculating. Then, the operation of ECODA is divided into 2 periods: initial data-collecting period and correlation clustering period. The correlation characteristics of the environment must be preserved in these periods.
- At initial data-collecting period: ECODA performs distance-based clustering to collect N_s samples serving for the entropy calculation process. The more the number of samples is, the more accurate the calculations of entropies and correlation coefficients are. The value of N_s will be decided before the deployment of the network. In our case, the value of Ns is 256. The initial data-collecting period is only implemented at the beginning of network operation and when we want to collect data of all nodes in the network to check correlation characteristics (in case of the changed correlation).
- After receiving enough N_s samples, BS begins to calculate entropy and entropy correlations coefficients. Based on the calculation results, the base station sends clustering information to sensor nodes to form clusters and then begins the correlation clustering period.
- At the correlation clustering period: the clusters are fixed (because the correlation regions are fixed). However, the cluster heads of the clusters are chosen to guarantee energy balance in the clusters and the connection paths from nodes in the clusters to their cluster head are established. Then the data transmission is done. In this period, data aggregation is done. The base station may use collected data to check the correlation characteristics and if the correlation has changed, the network will switch to the initial data-collecting period to form new correlation clusters. The longer the correlation clustering period is, the more advantages of correlation are exploited.
- Both the initial data-collecting period and correlation clustering period are separated into rounds. One round is comprised of a set-up phase and a steady-state phase as shown in Fig. 2. These phases will be explained in more detail later.
- In the set-up phase, the base station firstly determines the cluster formation and cluster head selection upon the initial data-collecting period or correlation clustering period. In this phase, the base station also specifies cluster heads with their members, shortest routing path from cluster members to their cluster heads (including intermediate nodes), active and nonactive nodes in a cluster (for representative aggregation).
- After the set-up phase, the data transmission path is established, and the network moves to the steady-state phase in which sensor nodes send data to their cluster heads and the cluster head sends data to the base station in a specified number of frames.

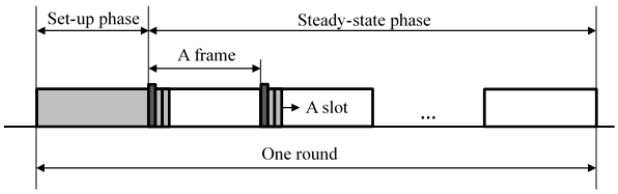

Fig. 2. Time scheduling for one round.

Set-up phase

In the initial data-collecting period, data is collected for discovering the correlation among nodes. This is done only at the beginning of network operation with full energy. The clustering process is performed by the base station with a possibility to use LEACH-C clustering [15]. However, LEACH-C is complicated, and because this period is implemented in a short period of time at the beginning of deploying the network, or when correlation relation has changed and we want to reform the cluster, we can simplify the implementation of this phase with distance-based clustering as follows.

In the set-up phase of the initial data-collecting period, at first, the base station determines the number of cluster *k*. According to[15], *k* can be chosen to be 5% of the total nodes in the network as the optimal value. Next, the sensed area is geographically divided into *k* equal parts. Each part corresponds to a cluster. After dividing the network into *k* clusters, the base station chooses the cluster head so that the total square of the distance from cluster head to cluster members is minimum. This selection ensures the minimization of the dissipated energy for transmitting data from cluster members to the cluster head.

After the initial data collection period, all nodes in the network are divided into correlation clusters. Then, in the set-up phase of the correlation clustering period, the selection of a cluster head for each cluster is done. Because the cluster head will receive data from all nodes in the same cluster, process and send them to the BS, it will dissipate more energy than the other nodes in the cluster. Therefore, the remaining energy should be considered when assigning the cluster head. The cluster head is chosen so that the total square of the distance from cluster head to cluster members is minimum, and the cluster head's remaining energy is larger than the average remaining energy of all nodes in the cluster. This minimizes the dissipated energy of data transmission from a cluster member to cluster head among available energy of nodes.

If representative aggregation is chosen, then, in the correlation clustering period, the selection of representative nodes is done by the base station in this phase. The choice of the representative node selection algorithm depends on the operation purpose. It is noted that with the purpose of maximizing information or coverage area, the representative nodes are usually fixed. Thus, it is difficult to obtain balance energy. In this paper, we choose the representative nodes to get balance energy. In a correlation cluster, nodes with the highest energy are chosen to be representative nodes.

In both periods, after determining the cluster head and cluster member, it is necessary to find the routing path from cluster members to their cluster head. There are various SPT (shortest path tree) algorithms to determine the optimal path from one point to another point. In this paper, an optimal routing algorithm in [16] is used because of its simplicity and energy efficiency. To establish the route from one node to its cluster head, this algorithm tries to choose the intermediate nodes that satisfy the following conditions:

- The intermediate node should have the maximum residual energy.
- The intermediate node should be as close to the cluster head as possible.
- The multi-hop path should be almost straight between the node and the cluster head.

Steady-state phase

After the set-up phase, the data transmission path is established, and the network moves to the steady-state phase in which sensor nodes send data to their cluster heads and the cluster head sends data to the base station. Because some nodes may not transmit data directly to the cluster head, but via intermediate nodes, the transmission process is the multi-hop type. In the steady-state phase, cluster head/intermediate nodes send a sending schedule to active nodes. With the initial data-collection period, nodes send data to intermediate/cluster head nodes and the intermediate/cluster head nodes forward the data to the next-hop/base station. With the correlation clustering period, the intermediate nodes collect sent data from other nodes, compress them with its sensed data and then send this compressed data to the upper intermediate nodes or cluster head. Cluster heads collect sensed data from their members, compress and send them to the base station.

Once the intermediate nodes/cluster head receives all the data, in the steady-state phase of the correlation clustering period, it can operate on the data such as performing data decompression/compression and then send to the base station. In this paper, Huffman based lossless compression [17] is used for compression. The fixed Huffman dictionary can be created by the BS based on distributions of sensed data. This dictionary is then sent to all nodes so that every node can encode and decode the compressed data easily. Later, the resultant data is sent to upper-level intermediate nodes (or cluster head/ base station).

5. Performance evaluation

Simulation setup

In order to obtain precise simulation results, the simulation model in [15] is used. The simulation model is implemented using MATLAB scripts.

The simulated network includes 400 sensor nodes uniformly distributed randomly in an area $[200m \times 200m]$. The simulation parameters are shown in Table 1. The distribution of network nodes is presented in Fig. 3.

Table 1. Simulation parameters

Parameter	Value
Sensing area $[m \times m]$	200×200
Base station position $[x, y]$	$[100, 275]$
Number of nodes	400
Initial energy $[J]$	0.5
The energy dissipated per bit $E_{elec}[J/bit]$	50×10^{-9}
Free space loss $\epsilon_{fs}[J/bit/m^2]$	10^{-11}
Multipath fading loss $\epsilon_{mp}[J/bit/m^4]$	1.3×10^{-15}
Aggregated energy $E_{DA}[J/bit]$	5×10^{-9}
Packet size $l[bit]$	4000
Correlation coefficient	0.6
Desired distortion	0.1

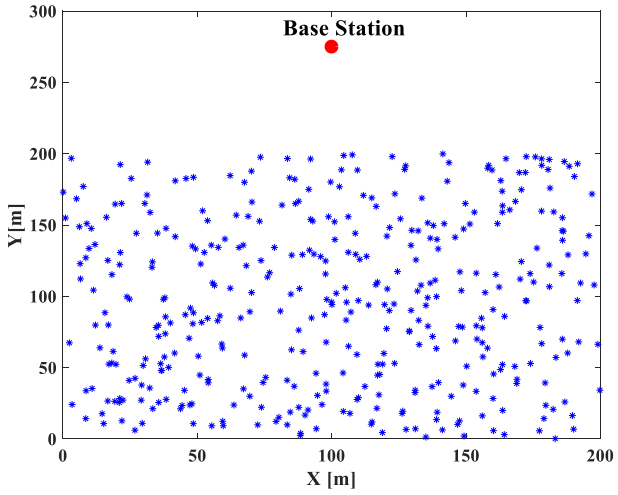

Fig. 3. Sensor node distribution in the 200mx200m sensing area.

a. The setting of the compression-based scheme

In this case, it is supposed that there are S correlation clusters. To simulate the situation of routing with SPT to the cluster head, in each cluster, the intermediate nodes are called group head. Nodes are in the same group if they are connected to the same intermediate node (or group head). Thus, a cluster is further divided into G groups. In each group, data is transmitted from group nodes to their group head. The group heads compress the obtained data and send it to the base station.

In the situation of compression at cluster head only, in each cluster, again, nodes are further divided into G groups. But it is not the same as the previous situation, the group head only transfers the data to cluster head without any compression. The cluster head received all data, compressed and sent to the base station. The routing path is illustrated as in Fig. 4.

b. The setting of the representative-based scheme

In this case, it is obvious that the lower the number of representative nodes is, the higher the energy conservation is. Therefore, two structures are chosen for simulation. In the first structure, there are 40 correlation clusters, i.e. there are 10 nodes on average in each cluster. In this case, the percentage of the representative is very high, about 80% of total nodes (with entropy correlation coefficient in this simulation is 0.6). In the second structure, the number of clusters is 16 (4% of the total number of nodes usually is chosen with distance-based energy-efficient routing protocols). In both structures, nodes transmit their data to their cluster head and cluster heads transmit all obtained data to the base station with or without compression.

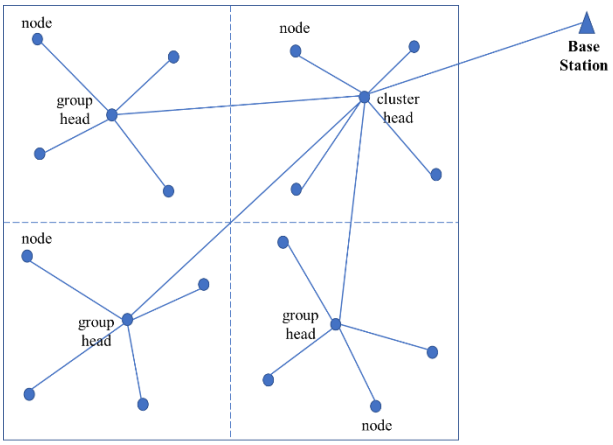

Fig. 4. Routing path of compression-based routing protocol.

Simulation results and discussions

To analyze results, we will evaluate the network lifetime, i.e. operational time of the network from the beginning until all nodes die). The network lifetime is calculated by several rounds. In addition, some other parameters are also considered such as the moment at which the first node/ half of the total number of nodes dies and half of the total energy remains.

a. The compression along SPT to the cluster head (CH)

- We consider three simulation situations. In the first situation, the network includes 16 correlation clusters ($S = 16$). In each cluster, nodes transfer data directly to their cluster heads, i.e. the whole cluster is a group ($G = 1$).
- In the second situation, the network includes 8 correlation clusters ($S = 8$), each cluster is divided into 2 groups ($G = 2$).
- In the third situation, the network includes 4 correlation clusters ($S = 4$), each cluster is divided into 4 groups ($G = 4$).

The second and third situations are full compression along SPT to the cluster head types. Fig. 5, Fig. 6 and Table 2 show the simulation results about total energy and a live node's number of the networks for 1200 rounds. It is found that the smaller the number of correlation clusters is, the better the performance of the network is.

From the first situation to the third situation, the network lifetime increases from 987 rounds to 1019 rounds and then 1142 rounds. The time that network lost half of its total energy increased from 197 rounds (20%

of the lifetime) to 353 rounds (34% of the lifetime) and then 438 rounds (38% of the lifetime), as shown in Fig. 5

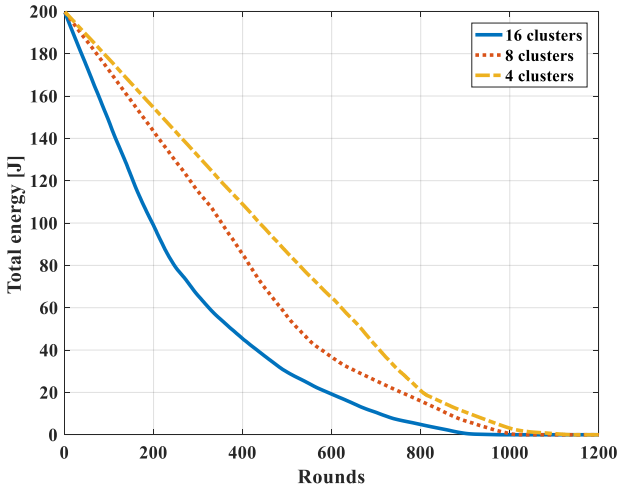

Fig. 5. The total energy in each round in case of compression along SPT to the CH.

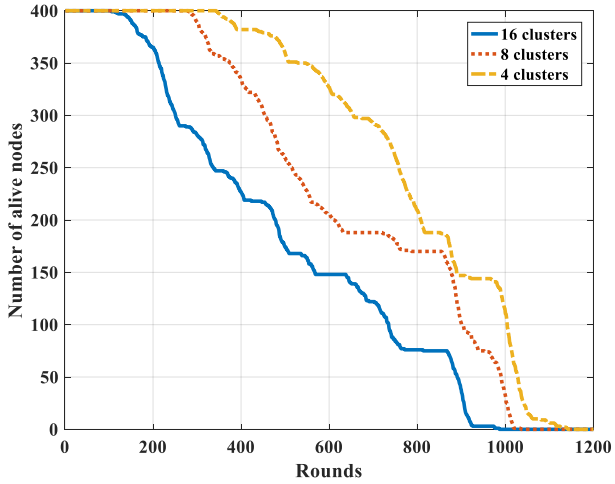

Fig. 6. Number of alive nodes in each round in case of compression along SPT to the CH.

Table 2. Simulation results in case of compression along SPT to the CH

Type	Network life (rounds)	The first node dies (rounds)	Half of the initial energy lost (rounds)	Half of the number of nodes die (rounds)
16 clusters	987	105	197	478
8 clusters	1019	280	353	613
4 clusters	1142	344	438	808

From Fig. 6, it can be found that the moment at which the first node died also increases from 105 rounds (10% of the lifetime) to 280 rounds (27% of the lifetime) and then 344 rounds (30% of the lifetime). The moment at which network lost half of the number of nodes also increases from 478 rounds (48% of the lifetime) to 613 rounds (60% of the lifetime) and then 808 rounds (71% of the lifetime). The reason is that the smaller the number of clusters is, the larger the number of nodes in each cluster is. The compression is more efficient with a high number of correlation data. Additionally, the smaller the number of clusters is, the smaller the dissipated energy of transmitting data to the base station far from nodes is.

In addition, for the second and third situations, the moment at which the network lost half of the total energy is quite close to the moment at which the first node dies. The network lost it's a half number of nodes after 60% and 71% of its lifetime. It means that the dissipated energy quite balances among nodes. The total energy reduces linearly until half of the number of nodes died. Then the speed of reduced energy is slow down. The reason is that the nodes far from the base station died, only nodes that are close to the base station still alive, thus the dissipated energy is reduced.

b. Representative aggregation

In this case, Fig. 7, Fig. 8 and Table 3 show the simulation results (total energy and the number of alive nodes) of the networks for 5000 rounds in the cases of 16 and 40 correlation clusters. In the case of 16 clusters, the network lifetime is 3805 rounds, while the network lifetime is only 1873 rounds for 40 clusters. The moment at which the first node died is 326 rounds (8.6% of the lifetime) in the case of 16 clusters and 15 rounds (0.8% of the lifetime) in the case of 40 clusters.

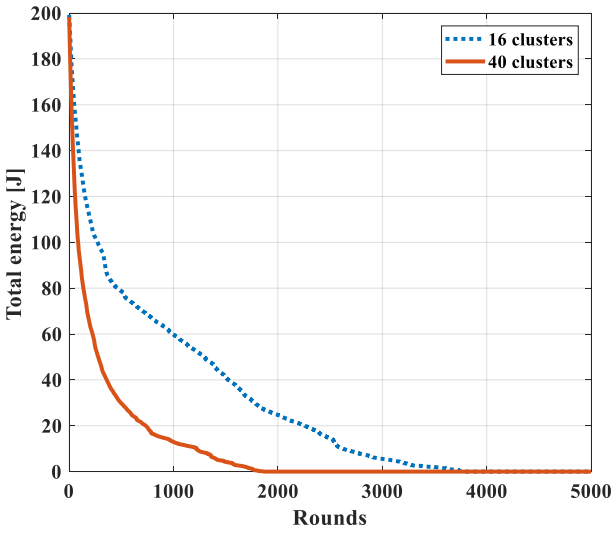

Fig. 7. The total energy in each round in the case of representative aggregation with 16 correlation clusters.

The moment at which network lost half of the total energy is 272 rounds (7.1% of the lifetime) in the case of 16 clusters and 87 rounds (4.6% of the lifetime) in the case of 40 clusters. The moment at which the network lost half of the number of nodes is 2589 rounds (68% of the lifetime) and 417 rounds (22% of the lifetime). From the simulation results, it is found that the smaller the

number of correlation clusters is, the better the performance of the network is.

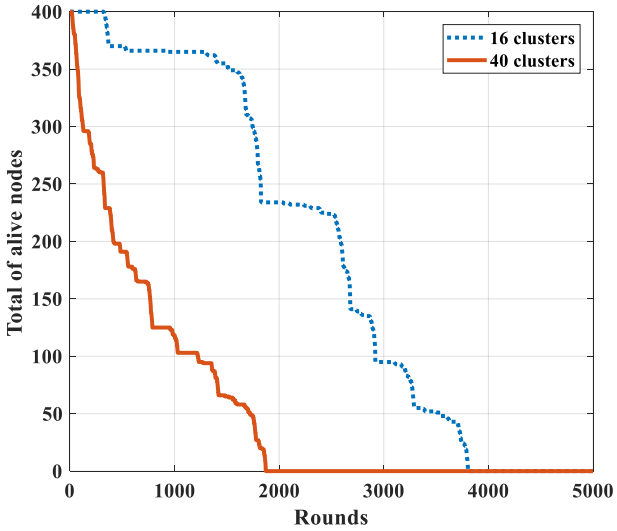

Fig. 8. Number of alive nodes in each round in the case of representative aggregation with 16 correlation clusters.

Table 3. Simulation results in the case of representative aggregation

Type	Network life (rounds)	The first node dies (rounds)	Half of the initial energy lost (rounds)	Half of the number of nodes die (rounds)
40 clusters	1873	15	87	417
16 clusters	3805	326	272	2589

Evaluation and comparison

To evaluate the effectiveness of ECODA, a distance-based energy efficiency protocol as LEACH-C is used for comparison in the following sections. In ECODA, two aggregation schemes including compression and representative aggregations can be used for this comparison purpose. Thus, we will consider two cases: ECODA with compression aggregation and ECODA with representative aggregation.

a. The case of ECODA with compression aggregation

In the case of 16 cluster heads, the shortest path from a node to its cluster head is the direct connection between them. Thus, compression is done only at the cluster head. Fig. 9 and Fig. 10 show the performance comparison between two protocols including total energy and number of live nodes.

It is found that ECODA is better than LEACH-C, except the lifetime parameters. The reason is that with compression, the dissipated energy is reduced much more than distance-based optimization.

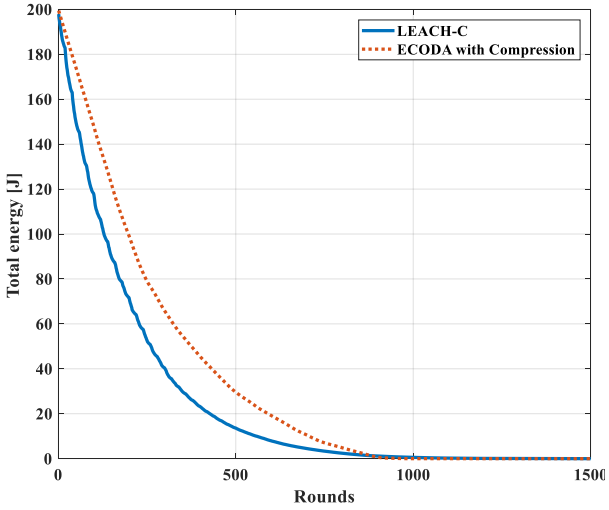

Fig. 9. Total energy comparison between LEACH-C and ECODA with compression aggregation in the case of 16 correlation clusters.

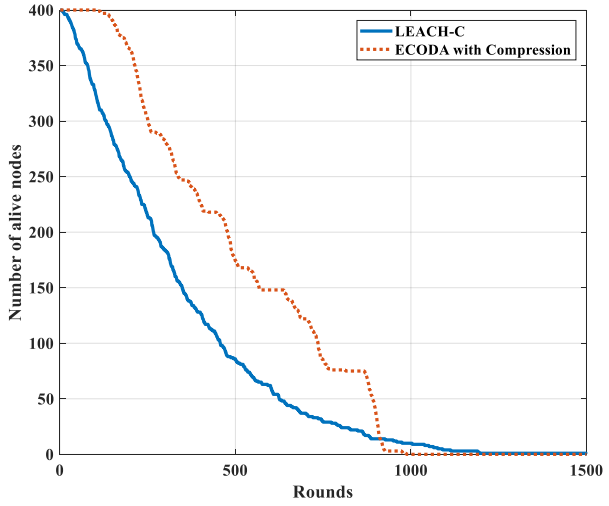

Fig. 10. Total energy comparison between LEACH-C and ECODA with compression aggregation in the case of 16 correlation clusters.

Table 4. Comparison between LEACH-C and ECODA with compression aggregation in the case of 16 correlation clusters

Type	Network life (rounds)	The first node dies (rounds)	Half of the initial energy lost (rounds)	Half of the number of nodes die (rounds)
LEACH-C	1500	5	128	267
ECODA	987	105	197	478

However, when the number of nodes is reduced, the

effectiveness of compression is decreased and if the number of nodes is small enough, there is a little difference between two cases: with and without compression. From that point, LEACH-C is better than ECODA.

In the case of 8 clusters (shown in Fig. 11 and Fig. 12 and Table 5), the performances of ECODA are also better than LEACH-C, especially in the case of the compression along SPT to cluster head (ECODA-SPT).

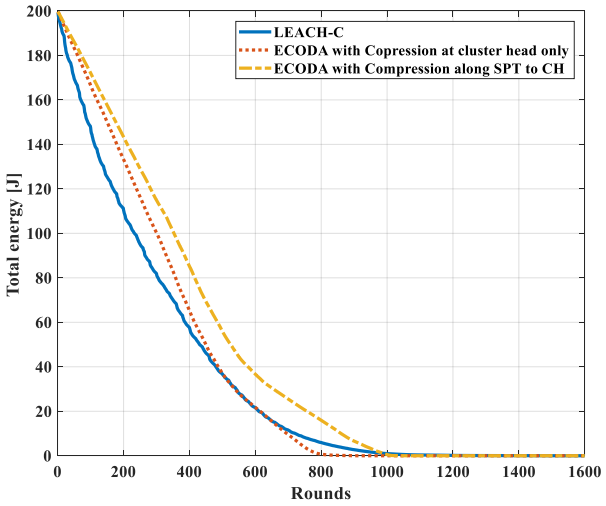

Fig. 11. Total energy comparison between LEACH-C and ECODA with compression aggregation in the case of 8 correlation clusters.

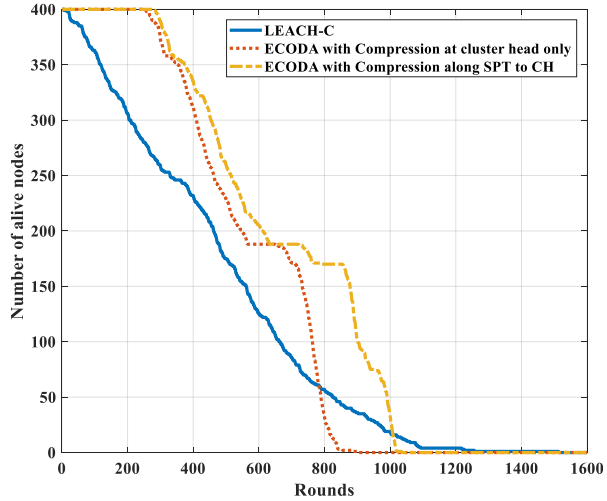

Fig. 12. Total energy comparison between LEACH-C and ECODA with compression aggregation in the case of 8 correlation clusters.

In the case of compression along SPT to cluster head, the compression was done along the way to the cluster head and at the cluster head. Therefore, the performance is much better. Only at the end of the network lifetime, when the number of alive nodes becomes small, the compression does not affect much of the dissipated energy, then the distance-based optimization will be better.

Table 5. Comparison between LEACH-C and ECODA in the case of 8 correlation clusters

Type	Network life (rounds)	The first node dies (rounds)	Half of the initial energy lost (rounds)	Half of the number of nodes die (rounds)
LEACH-C	1514	5	233	462
ECODA	900	250	301	549
ECODA-SPT	1019	280	353	613

As same as the case of 16 cluster heads, the lifetime of the distance-based protocol is longer than ECODA but the other parameters are not better than ECODA. In the case of compression at cluster head only, the compression is done only at the cluster head, thus the compression effect is not clear in the case of a small number of the cluster head.

b. The case of ECODA with representative aggregation

Fig. 13, Fig. 14 and Table 6 show the performance comparison between LEACH-C and ECODA with representative aggregation (with and without compression) in the case of 16 cluster heads.

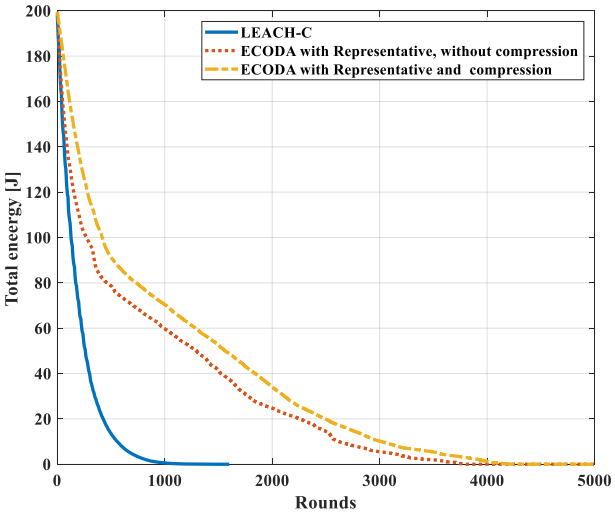

Fig. 13. Total energy comparison between LEACH-C and ECODA with representative aggregation in the case of 16 correlation clusters.

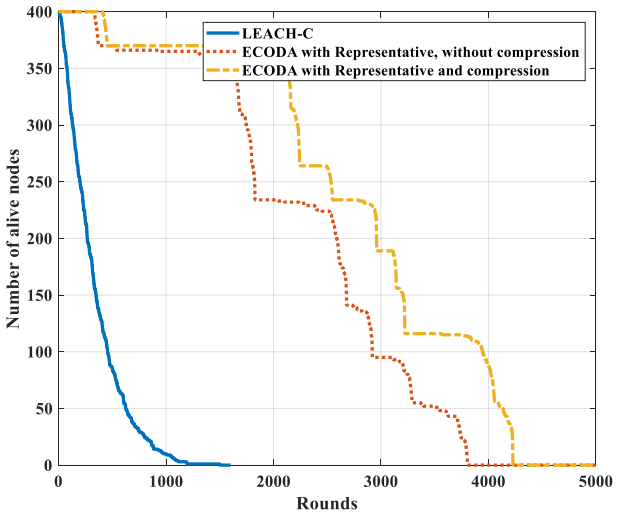

Fig. 14. Number of a live nodes comparison between LEACH-C and ECODA with representative aggregation in the case of 16 correlation clusters.

Because distortion is accepted, in ECODA, some nodes can sleep to save energy; therefore the performance is much better than the distance-based protocol. If the compression is done in the cluster heads, then the performance is much better in ECODA.

It is noted that in all considered routing protocols, the total energy is reduced quite fast at the beginning and then reduced slower. The reason is that at the beginning area, nodes at all sensing area dissipate energy to send data to the base station. Then the nodes far from the base station died, the remained alive nodes are close to the base station, thus the dissipated energy is smaller.

Table 6. Comparison between LEACH-C and ECODA with representative aggregation in the case of 16 correlation clusters

Type	Network life (rounds)	The first node dies (rounds)	Half of the initial energy lost (rounds)	Half of the number of nodes die (rounds)
LEACH-C	1500	5	128	267
ECODA-Rep-without compression	3805	326	272	2589
ECODA-Rep- with compression	4228	409	419	2958

6. CONCLUSIONS

In this paper, we have proposed an **E**ntropy **CO**rrelation clustering for **D**ata **A**ggregation (ECODA) protocol for a wireless sensor network in correlation environments. The operation process of the protocol is divided into two periods including initial data-collecting period and correlation clustering period. The initial data-collecting period is at the beginning of the operation process in order to get the data for correlation identification. The next period is the main process where the network is with correlation clustering and deploys the proposed clustering and data aggregation schemes. In each period, the base station implements clustering and establishes the connection among networks in the setup phase. Then, in the steady-state phase, the data is sent to the station. The base station always uses received data to re-identify the correlation among nodes in the network.

In addition, simulation has been done with various conditions and ECODA is compared with LEACH-C protocol. It is shown that ECODA has better performance with better energy balance. The simulation results validated the effectiveness of ECODA.

In the future, ECODA will be implemented in a real network with correlation characteristics. In addition, the development of Distributed Source Coding which is the most efficient compression scheme for ECODA will be considered. Moreover, the combination of the proposed spatial correlation model with a temporal correlation of measured data will be considered to further improve the energy efficiency of ECODA.

REFERENCES

[1] I. F. Akyildiz, et. al. (2002). Wireless Sensor Networks: A Survey. Computer Networks (Elsevier) Journal, vol. 38, no. 4, (March 2002), 393-422.

[2] Srisooksai, et al. (2012). Practical data compression in wireless sensor networks: A survey. Journal of Network and Computer Applications 35.1, 37-59.

[3] M. C. Vuran, et. al. (2004). Spatio-temporal correlation: Theory and applications for wireless sensor networks, Computer Network, vol. 45, no. 3, pp. 245–259.

[4] Shakya, Rajeev K., et. al. (2013). Generic correlation model for wireless sensor network applications. IET Wireless Sensor Systems 3.4 (2013): 266-276.

[5] Liu, Chong, et. al. (2007). An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation" IEEE Trans. on parallel and distributed systems 18.7.

[6] Gupta, Himanshu, et al. (2008). Efficient gathering of correlated data in sensor networks. ACM Tran. on Sensor Networks (TOSN)4.1.

[7] Ma, Yajie, et al. (2011). Distributed clustering-based aggregation algorithm for spatial correlated sensor networks. IEEE Sensors Journal 11.3 (2011): 641-648.

[8] Yuan, Fei, et. al. (2014). Data density correlation degree clustering method for data aggregation in WSN. IEEE Sensors Journal 14.4: 1089-1098.

[9] Pattem, et. al. (2008). The impact of spatial correlation on routing with compression in wireless sensor networks. ACM Trans. on Sensor Networks (TOSN) 4.4: 24.

[10] Dai, Rui, and et. al. (2009). A spatial correlation model for visual information in wireless multimedia sensor networks. IEEE Trans. on Multi. 11.6 (2009): 1148-1159.

[11] Wang, Fan, et al. (2016). Energy-efficient clustering using correlation and random update based on data change rate for wireless sensor networks. IEEE Sensors Journal (2016): 1-1.

[12] D. Maeda, et.al. (2007). Efficient Clustering Scheme Considering Non-uniform Correlation Distribution for Ubiquitous Sensor Networks. IEICE Trans. on Fund. of Electronics, Comm. and Computer Sciences E90-A (7), (2007), 1344-1352.

[13] Nguyen Thi Thanh Nga, et. al. (2018). Entropy Correlation and Its Impacts on Data Aggregation in a Wireless Sensor Network. Sensors (2018); 18(9):3118

[14] Nguyen Thi Thanh Nga et. al. (2018). Entropy Correlation-based Clustering Method for Representative Data Aggregation in Wireless Sensor Networks. International Journal of Sensor Networks, 2018, Vol.28 No.4, pp.270 - 283

[15] Heinzelman W. B., Chandrakasan A. P., and Balakrishnan H., (2002). An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun., vol. 1, no. 4, pp. 660–670.

[16] Tran-Quang V., and Miyoshi T., (2008). Adaptive Routing Protocol with Energy Efficiency and Event Clustering for Wireless Sensor Networks. IEICE Trans. Commun., vol. E91–B, no. 9, pp. 2795–2805.

[17] Medeiros H. P., Maciel M. C., Demo Souza R., and Pellenz M. E., (2014) "Lightweight data compression in wireless sensor networks using Huffman coding," Int. J. Distrib. Sens. Networks, 10(1), 672921.