An Analysis of Variance-based Methods for Data Aggregation in Periodic Sensor Networks

Hassan Harb[1,3,*], Abdallah Makhoul[1], David Laiymani[1], Oussama Bazzi[2], and Ali Jaber[3]

[1] FEMTO-ST Laboratory, DISC department, University of Franche-Comté, Belfort, France
hassan.moustafa_harb@univ-fcomte.fr
{abdallah.makhoul, david.laiymani}@univ-fcomte.fr
[2] Department of Physics and Electronics, Lebanese University, Beirut, Lebanon
obazzi@ul.edu.lb
[3] Department of Computer Science, Lebanese University, Beirut, Lebanon
ali.jaber@ul.edu.lb

Abstract. Given the vast area to be covered and the random deployment of the sensors, wireless sensor networks (WSNs) require scalable architecture and management strategies. In addition, sensors are usually powered by small batteries which are not always practical to recharge or replace. Hence, designing an efficient architecture and data management strategy for the sensor network are important to extend its lifetime. In this paper, we propose energy efficient two-level data aggregation technique based on clustering architecture with which data is sent periodically from nodes to their appropriate Cluster-Heads (CHs). The first level of data aggregation is applied at the node itself to eliminate redundancy from the collected raw data while the CH searches, at the second level, nodes that generate redundant data sets based on the variance study with three different Anova tests. Our proposed approach is validated via experiments on real sensor data and comparison with other existing data aggregation techniques.

Keywords: periodic sensor networks (PSNs), data aggregation, clustering architecture, identical nodes behaviour, one way Anova model.

1 Introduction

Wireless Sensor Networks (WSNs) have become one of the innovative technologies that are widely used nowadays. One of the advantages of these networks is their ability to operate unattended in harsh environments in which contemporary human-in-the-loop monitoring schemes are risky, inefficient and sometimes infeasible (see Abbasi, A. and Younis, M. [1]). With the capabilities of pervasive surveillance, WSN have attracted significant attention in many applications, such as habitat monitoring (see Rozyyev, A. et al. [2]), environment monitoring (see Sabri, N. et al. [3] and Aslan, Y.E. et al. [4]) and military surveillance (see

Qian, H. et al. [5] and Padmavathi, G. et al. [6]). In such networks, sensors are expected to be remotely deployed, e.g. via helicopter or clustered bombs, in a wide geographical area to monitor the changes in the environment and send back the collected data to a specific node called the "sink". Nevertheless, sensors in such environment are energy-constrained and their batteries cannot be replaced. Therefore, it is very important to limit the energy consumption of sensors in order to extend the network's lifetime as long as possible.

Due to a random and dense deployment, nodes may have overlapping sensing ranges, such that events can be detected by multiple sensor nodes providing a redundancy in sensed data. Moreover, since data transmission is more demanding than computational operations in terms of energy consumption, the volume of data transmitted must be minimized. This leads to the requirement of better data aggregation and data mining techniques. To that effect, data aggregation has been proved as an effective method to achieve power efficiency by reducing data redundancy and minimizing bandwidth usage (Di Pietro, R. et al. [7]) while data mining deals with extracting knowledge from large continuous arriving data from WSNs (Azhar, M. et al. [8]).

On the other side, clustering is considered as an efficient topology control method in WSN, which can increase network scalability and lifetime (Mirhadi, P. et al. [9]). With clustering, data collected by sensor nodes are processed at intermediate nodes, called Cluster-Heads (CHs), in order to eliminate redundancy and send only the useful information to the sink (Fig. 1). In this paper, we use the periodic data collection approach, in which each sensor node sends periodically (at each period p) its data to the appropriate CH. We propose an energy efficient two-level data aggregation technique which applies at each cluster separately. The first level is applied at the sensor node itself in order to eliminate redundancy from data collected by the sensor at each period p before sending them to their proper CH. Then, when the CH receives data from all its members (nodes) we propose to use the one way Anova model with three different tests (Fisher, Tukey and Bartlett) to detect nodes with identical behaviour which generate redundant data logs or sets. The aim is to reduce data redundancy generated by neighboring nodes based on the variance study in order to eliminate redundancy before sending final data to the sink.

The rest of this paper is organized as follows; Section 2 presents related work on data aggregation in the sensor networks. Section 3 describes the first phase of our technique which we called member node aggregation. In Section 4, we present the second level, called CH aggregation, which is based on one way Anova model. Experimental results are exposed in Section 5. Finally, we conclude our paper and we provide our directions for future work in Section 6.

2 Related Work

In WSN, many data aggregation studies have been made based on clustering schemes, such as DDCD proposed by Yuan, F. et al. [10] and DUCA proposed by Enam, R.N. et al. [11]. The main objective of these works is balancing and

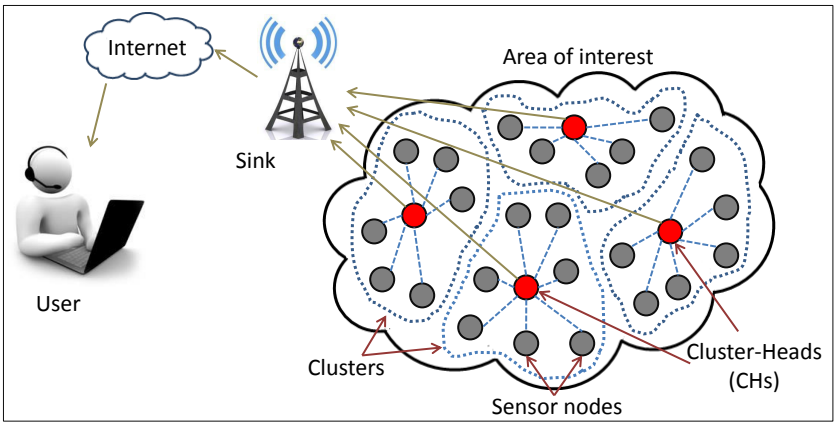


Fig. 1. Wireless sensor network based on two-tier single-hop clustering architecture.

reducing energy consumption over the whole network. In each cluster, the sensors communicate data to their CH that aggregates data and thus reduces the size of data to be transmitted to the sink. Recently, Tripathi, A. et al. [12] and Nokhanji, N. and Hanapi, Z.M. [13] present a comprehensive overview about different data aggregation techniques and clustering routing protocols proposed in the literature for WSNs.

Zou, P. and Liu, Y. [14] propose a Distributed K-mean Clustering (DKC) method for WSN. On the basis of DKC, the authors build a network data aggregation processing mechanism based on adaptive weighted allocation of WSN. DKC algorithm is mainly used to process the testing data of bottom nodes in order to reduce the data redundancy. Tran, K.T-M. and Oh, S.-H. [15] propose a data aggregation based clustering scheme for underwater wireless sensor networks (UWSNs) which involves four phases. The goals of these phases are to reduce the energy consumed in the overall network, increasing the throughput, and minimizing data redundancy. Kumar, S. et al. [16] propose a M-EECDA (Multihop Energy Efficient Clustering & Data Aggregation Protocol for Heterogeneous WSN). The protocol combines the idea of multihop communications and clustering for achieving the best performance in terms of network life and energy consumption. M-EECDA introduces a sleep state and three tier architecture for some cluster heads to save energy in the network.

Some other works in data aggregation are not based on clustering scheme: Chao, C.-M. and Hsiao, T.-Y. [17] propose a structure-free and energy-balanced data aggregation protocol, SFEB. SFEB features both efficient data gathering and balanced energy consumption, which result from its two-phase aggregation process and the dynamic aggregator selection mechanism. Li, G. and Wang, Y. [18] propose an automatic auto regressive-integrated moving average modeling-based data aggregation scheme in WSNs. The main idea behind this scheme is to decrease the number of transmitted data values between sensor nodes and

aggregators by using time series prediction model. Shan, M. et al. [19] study the problem of building maximum lifetime shortest path aggregation trees in WSNs. When the shortest path trees are built, the authors transformed the problem into a load balancing scheme at each level of the fat tree and solved it by a centralized approach in polynomial time. Shim, Y. and Kim, Y. [20] propose a data aggregation with multiple sinks in an Information-Centric Wireless Sensor Network with an ID-based information-centric network, in order to reduce the energy-transmission cost.

Bahi, J. et al. [21] study a new area within filtering aggregation problem, the Prefix-Frequency Filtering (PFF) technique. Further to a local processing at sensor node level, PFF uses Jaccard similarity function at aggregators level to identify similarities between near sensor nodes and integrate their sensed data into one record. Aiming to decrease data latency, Harb, H. et al. in [22] and [23] propose two optimizations of the PFF technique based on suffix filtering and k-means algorithm. Among all optimizations, PFF stays a hard technique for the aggregator in terms of data latency and energy consumption. In this paper, we adapt the same scenario as proposed by Bahi, J. et al. [21] while we propose a new technique. In the new technique, we propose a two-level data aggregation, the first one, at the node level, which we call member node aggregation in which each member node sends, at each period p, its aggregated set of data to the appropriate CH. At the second level, CH aggregates all the sets of data coming from its member nodes based on the variance between their measurements, before sending them to the sink.

3 First Level: Member Node Aggregation

In periodic sensor networks (PSNs), each sensor node i takes a new measurement y_{is} at each time slot s. Then node i forms a new vector of captured measurements $M_i = [y_{i_1}, y_{i_2}, \dots y_{i_\mathcal{T}}]$ at each period p, where $\mathcal{T}$ is the total number of measures taken at the period p, and sends it to the appropriate CH (see Bahi, J. et al. [21]). Fig. 2 shows an example of PSN where each sensor node takes one data measurement each ten minutes, e.g. $s = 10$ minutes, and send its set of collected data which contains six measures, e.g. $\mathcal{T}$=6, to the CH at the end of each hour.

Consequently, one of the important design considerations associated with the periodic sampling data model is that the dynamics of the monitored conditions can slow down or speed up (Laiymani, D. and Makhoul, A. [24]). Thus, it is likely that a sensor node takes the same (or very similar) measurements several times, especially when s is too short, which make the sensor node forwards more redundant data to the CH during each period. In this phase of aggregation, which called member node aggregation, we allow each sensor node to identify and remove duplicate data measurements among data collected in each period in order to reduce the size of the set M_i before sending it to the CH. In order to identify the similarity between two measures, we provide the two following definitions:

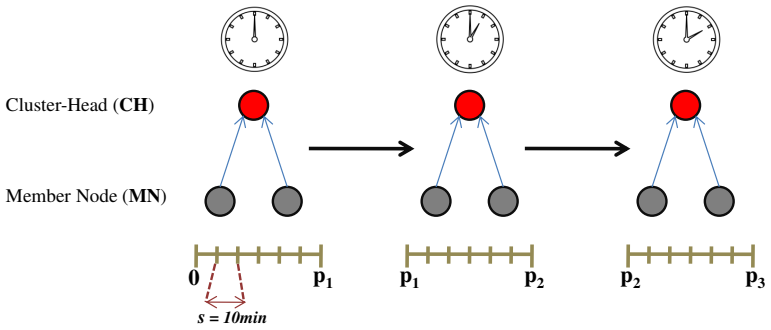


Fig. 2. Illustrative example of periodic sensor network (PSN).

Definition 1 (*Similar* function). *We define the Similar function between two measurements as:*

$$Similar(y_i, y_j) = \begin{cases} 1 & \text{if} \quad \|y_i, y_j\| \leq \delta, \\ 0 & \text{otherwise.} \end{cases}$$

where δ is a threshold determined by the application. Furthermore, two measures are similar if and only if their *Similar* function is equal to 1.

Definition 2 (Measure's weight, $wgt(y_i)$). *The weight of a measurement y_i is defined as the frequency of the same or similar (according to the Similar function) measurements in the same set.*

For each new sensed measurement (at each slot s), a sensor node i searches for the similar measure already captured in the same period p. If a similar measurement is found, the sensor deletes the new measure while incrementing the weight of the existing measure by one, else, the sensor adds the new measure to the set and initializes its weight to 1. For more details about this algorithm see Bahi, J. et al. [21].

Based on the above definitions, we provide two other definitions:

Definition 3 (Cardinality of the set M_i, $|M_i|$). *The cardinality of the set M_i is equal to the number of elements in M_i.*

Definition 4 (Weighted Cardinality of the set M_i, $Card_w(M_i)$). *The weighted cardinality of the set M_i is equal to the sum of all measures' weights in M_i as follow: $Card_w(M_i) = \sum_{k=1}^{|M_i|} wgt(m_k)$, where $m_k \in M_i$.*

In this paper, we consider that all sensor nodes operate at the same sampling rate, and every node captures τ measures in each period p. Thus we can deduce that for every received set M_i from node i we have: $Card_w(M_i) = \tau$.

At the end of each period p, each member node i will possess a set of reduced measures associated to their corresponding weight. The second step is to send it to the appropriate CH which in its turn aggregates the data sets coming from different member nodes.

4 Second Level: CH Aggregation

At this level of aggregation, each CH receives all the sets of measurements with their weights sent from its member nodes, at the end of each period. The idea is to identify all pairs of member nodes that generate redundant sets in order to eliminate duplication before sending them to the sink. Therefore, one way Anova model is an effective technique that can determine duplicated sets based on the variance between their measures. The Anova produces an F-statistic, the ratio of the variance calculated based on the measurements in the sets. F can be calculated in different manners depending on the statistic tests proposed in the Anova model. The sets are considered duplicated if the calculated F is less than the critical value of the F-distribution (or F_s) for some desired false-rejection probability (risk α). Laiymani, D. and Makhoul, A. [24] used one way Anova model and Fisher test in PSN at the level of node member to adapt its sampling rate. In this paper, we use the one way Anova model at the CH level, while comparing three different tests (Fisher, Tukey and Bartlett) in order to identify identical nodes behaviour.

4.1 One-Way ANalysis Of VAriance: ANOVA

In this part, we present a statistical model to study the variance between measurements in the data sets in order to find all pairs of member node that generate redundant data. Therefore, one-way Anova is used to find out if the means of data sets are significantly different or if they are relatively the same. In PSN, we assume that each sensor node takes τ measures of temperature or humidity within a period p.

When receiving data sets coming from its member nodes at each period, CH computes the variation between every pair of sets. Therefore, it uses the one way Anova to test whether or not the means of every pair are equal. In case that a pair of sets notices low differences variance, CH considers that the two member nodes generate redundant data. After identifying all pairs of redundant sets, CH uses selecting sets algorithm proposed in the later subsection to select final sets to be sent to the sink, while conserving the integrity of information.

We suppose that measures generated by each member node i at each period p are independent, then we denote by $\overline{Y_i}$ and σ_i^2 the mean and the variance of the set M_i generated by the member node i, and by $\overline{Y}$ the mean of the pair of sets (M_i, M_j) generated by the member node i and j respectively as follows:

$$\overline{Y_i} = \frac{1}{Card_w(M_i)} \sum_{k=1}^{|M_i|} \big(y_{ik} \times wgt(y_{ik})\big),\ \sigma_i^2 = \frac{1}{Card_w(M_i)} \sum_{k=1}^{|M_i|} \big(wgt(y_{ik}) \times (y_{ik} - \overline{Y_i})^2\big),$$

$$\overline{Y} = \frac{1}{Card_w(M_i)} \sum_{k=1}^{|M_i|} \big(y_{ik} \times wgt(y_{ik})\big) + \frac{1}{Card_w(M_j)} \sum_{k=1}^{|M_j|} \big(y_{jk} \times wgt(y_{jk})\big),$$

where $y_{ik} \in M_i$ and $y_{jk} \in M_j$.

Since $Card_w(M_i) = Card_w(M_j) = \mathcal{T}$:

$$\overline{Y_i} = \frac{1}{\mathcal{T}} \sum_{k=1}^{|M_i|} \big(y_{ik} \times wgt(y_{ik})\big),\ \sigma_i^2 = \frac{1}{\mathcal{T}} \sum_{k=1}^{|M_i|} \big(wgt(y_{ik}) \times (y_{ik} - \overline{Y_i})^2\big),$$
$$\overline{Y} = \frac{1}{2 \times \mathcal{T}} \Big(\sum_{k=1}^{|M_i|} \big(y_{ik} \times wgt(y_{ik})\big) + \sum_{k=1}^{|M_j|} \big(y_{jk} \times wgt(y_{jk})\big) \Big),$$

where $y_{ik} \in M_i$ and $y_{jk} \in M_j$.

The total variation (ST), in a pair of sets, is the sum of the variation (SR) within each set and the variation (SF) between the sets. SF represents what is often called "explained variance" or "systematic variance". We can think of this as the variance that is due to the independent variable, the difference among the two sets. For example the difference between measures in two or more different sets. SR represents what is often called "error variance". This is the variance within sets, variance that is not due to the independent variable. For example, the difference between measures in the same set. The whole idea behind the analysis of variance, in a pair of sets, is to compare the ratio of the variance between the sets to the variance within each set in this pair. If the variance caused by the interaction between the measures, in a pair of sets, is much larger than the variance that appears within the sets, then it is because the means arent the same. Let us consider:

$$ST = SR + SF \Rightarrow$$

$$\sum_{l}^{\{i,j\}} \sum_{k=1}^{|M_l|} \big(wgt(y_{lk}) \times (y_{lk} - \overline{Y})^2\big) = \sum_{l}^{\{i,j\}} \sum_{k=1}^{|M_l|} \big(wgt(y_{lk}) \times (y_{lk} - \overline{Y_l})\big)^2 + \mathcal{T} \sum_{l}^{\{i,j\}} (\overline{Y_l} - \overline{Y})^2 \tag{1}$$

4.2 Mean's Period Verification

In this section, we use three tests in the Anova model (Fisher, Tukey and Bartlett), to compute the means and the variances for every pair of sets, then to decide if the sets in this pair are redundant or not.

4.2.1 Fisher Test

The Fisher's test or F-test is a statistical hypothesis test for testing the equality of two variances by taking the ratio of the two variances and ensuring that this ratio does not exceed a certain theoretical value (find in Fisher's table). In the case of PSN, we compare, in a pair of sets, the ratio of the variance between the sets (SF) to that within each set in this pair (SR).

The general formula for the F-test is:

$$F = \frac{SF/(J-1)}{SR/(N-J)}$$

where J is the number of compared sets and N is the number of total measures in the compared sets. Therefore, J is equal to 2 in our case while N is equal to $2 \times \mathcal{T}$ (because $Card_w(M_i) = Card_w(M_j) = \mathcal{T}$).

Then, we deduce:

$$F = 2 \times (\mathcal{T} - 1) \times \frac{SF}{SR} \tag{2}$$

For each pair of sets, the CH will test the hypothesis that means of sets are the same or not. If the hypothesis is correct then, F will have a Fisher distribution, with $F(1, 2\times(\mathcal{T}-1))$ degrees of freedom. The hypothesis is rejected if the F calculated from the measures is greater than the critical value of the F distribution for some desired false-rejection probability (risk α). Let $F_t = F_{1-\alpha}(1, 2 \times (\mathcal{T}-1))$.

The decision is based on F and F_t :

- if $F > F_t$ the hypothesis is rejected with false-rejection probability α, and the variance between the sets are significative.
- if $F \leq F_t$ the hypothesis is accepted.

4.2.2 Tukey Test

The Tukey's post-hoc test, proposed by Hall, R. [25], is a single-step multiple comparison procedure and statistical test. It can be used to calculate the difference between the means of two or multiple sets. Tukey's test works by defining a value known as Honest Significant Difference (HSD). HSD represents the minimum distance between the means of two sets to be considered statistically significant.

Tukey's test can be applied to a pair of sets (M_i, M_j) based on the following equations:

$$SS_{total} = \sum_{l}^{\{i,j\}} \sum_{k=1}^{|M_l|} \left(wgt(y_{lk}) \times y_{lk}^2\right) - \frac{\left(\sum_{l}^{\{i,j\}} \sum_{k=1}^{|M_l|} \left(wgt(y_{lk}) \times y_{lk}\right)\right)^2}{2 \times \mathcal{T}} \tag{3}$$

$$SS_{among} = \frac{\left(\sum_{k=1}^{|M_i|} \left(wgt(y_{ik}) \times y_{ik}\right)\right)^2 + \left(\sum_{k=1}^{|M_j|} \left(wgt(y_{jk}) \times y_{jk}\right)\right)^2}{\mathcal{T}} - \frac{\left(\sum_{l}^{\{i,j\}} \sum_{k=1}^{|M_l|} \left(wgt(y_{lk}) \times y_{lk}\right)\right)^2}{2 \times \mathcal{T}} \tag{4}$$

$$SS_{within} = SS_{total} - SS_{among}; \quad df_{among} = 1; \quad df_{within} = 2 \times \tau - 2$$

$$MS_{among} = \frac{SS_{among}}{df_{among}}; \; MS_{within} = \frac{SS_{within}}{df_{within}}; \; F = \frac{MS_{among}}{MS_{within}}$$

Where:

- SS_{within} : Sum of squares within the pair of sets (M_i, M_j),
- SS_{among} : Sum of squares between the sets in the pair (M_i, M_j),
- MS_{within} : Mean squares within the pair of sets (M_i, M_j),
- SS_{among} : Sum of squares between the sets in the pair (M_i, M_j).

Therefore, when we calculate F we check to see if it is statistically significant based on studentized range distribution table with appropriate degrees of freedom $F_t = df(df_{among}, df_{within})$. The decision is based on F and F_t:

- if $F > F_t$ the hypothesis is rejected with false-rejection probability α, and the variance between the sets M_i and M_j are significative.
- if $F \leq F_t$ the hypothesis is accepted.

4.2.3 Bartlett Test

The Bartlett's test [26] is used to test if two or multiple data sets are from populations with equal variances. Equal variances across data sets is called homogeneity of variances. Some statistical tests, for example the analysis of variance, assume that variances are equal across data sets. The Bartlett test can be used to verify that assumption. Bartlett's test is used to test the null hypothesis, H_0 that variances of all data sets are equal against the alternative that at least two are different. In our case, we test the hypothesis H_0 for every pair of sets (M_i, M_j) each having a size τ and with variances σ_i^2 and σ_j^2 respectively. Bartlett's test statistic is:

$$F = \frac{2 \times (\tau - 1) \ln(\sigma_p^2) - (\tau - 1)(\ln \sigma_i^2 + \ln \sigma_j^2)}{\lambda} \tag{5}$$

where :

$$\lambda = 1 + \frac{1}{2 \times (\tau - 1)} \tag{6}$$

and σ_p^2 is the pooled variance, which is a weighted average of the period variances and it is defined as:

$$\sigma_p^2 = \frac{1}{2 \times (\tau - 1)} \times (\sigma_i^2 + \sigma_j^2)$$

Bartlett's test has approximately a $(J-1)$ degrees of freedom where J is equal to 2 in our case. Thus the null hypothesis is rejected if $F > T_{J-1,\alpha}$ (where $T_{J-1,\alpha}$ is the upper tail critical value for the T_{J-1} distribution). We suppose that $F_t = T_{J-1,\alpha}$, thus the decision is based on the following rule:

- if $T > F_t$ the hypothesis is rejected with false-rejection probability α, and the variance between the sets M_i and M_j are significative.
- if $T \leq F_t$ the hypothesis is accepted.

4.3 Aggregation at the CH level

In this section, we present the algorithms that follow each CH to find redundant data sets based on Anova model, then to remove redundancy before sending them to the sink.

4.3.1 Sets redundancy searching

In our technique, one way Anova model is used to find all pairs of sets that have low variance between their measures. Algorithm 1 describes how these pairs are found in our technique. For every pair of sets (M_i,M_j), we calculate the corresponding F score as described in each test presented before (line 4). Then, we search the corresponding threshold F_t based on the probability table for each test with the appropriate degrees of freedom (line 5). Finally, we conclude that M_i and M_j are redundant sets in the case where the variance between their measures (F) is less than the threshold F_t (line 6).

Algorithm 1 CH aggregation algorithm.

Require: Set of measures' sets $M = \{M_1, M_2 ... M_n\}$.
Ensure: All pairs of sets (M_i, M_j), such that $F \leq F_t$.
1: $S \leftarrow \emptyset$
2: **for** each set $M_i \in M$ **do**
3: **for** each set $M_j \in M$ such that $M_j \neq M_i$ **do**
4: compute F for (M_i, M_j)
5: find F_t
6: **if** $F \leq F_t$ **then**
7: $S \leftarrow S \cup \{(M_i, M_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: return S

4.3.2 Redundant sets reduction

After identifying all pairs of redundant sets, the CH deletes redundant data sets sent from neighboring sensors in order to reduce the amount of data transmitted to the sink while conserving the integrity of information. Algorithm 2 shows how the CH selects the data sets to be sent to the sink among the pairs of redundant received sets. For each similar pair of set, the CH chooses the one having the highest cardinality (line 3), then it sorts it in increasing order of the measures to accelerate a measure search[4]. After that, for each measure in the other set, CH searches for its similar in the highest set and merges its weight to the similar one found (line 9). Otherwise, CH adds the measure with its weight to the highest set (line 11). The objective of merging the weights of similar measures is to save the information without any loss. Finally, the CH removes all pairs of redundant sets that contain M_i or M_j from the set of pairs (which means it will not check them again) (line 15).

Algorithm 2 selecting sets algorithm.

Require: All pairs of sets (M_i, M_j), such that $F \leq F_t$.
Ensure: List of selected sets, L.
1: $L \leftarrow \emptyset$
2: **for** each pair of sets (M_i, M_j) **do**
3: Consider$|M_i| \geq |M_j|$
4: $M_i \leftarrow \text{sort}(M_i, |M_i|)$, M_i is sorted in increasing order of the measures
5: **for** $k = 1 \rightarrow |M_j|$ **do**
6: Search similar of $M_j[k]$ in M_i
7: find $M_i[l]/\ Similar(M_j[k], M_i[l]) = 1$
8: **if** $M_i[l]$ exists **then**
9: $wgt(M_i[l]) \leftarrow wgt(M_i[l]) + wgt(M_j[k])$
10: **else**
11: $M_i \leftarrow M_i \cup \{(M_j[k], wgt(M_j[k]))\}$
12: **end if**
13: **end for**
14: $L \leftarrow L \cup \{M_i\}$
15: Remove all pairs of sets containing one of the two sets M_i and M_j
16: **end for**

5 Performance Evaluation

In this section, we present the experimental results which evaluate the performance of our proposed technique. The objective of these experiments is to confirm that our technique can successfully achieve desirable results for energy conservation in PSNs. Therefore, we used the publicly available Intel Lab dataset which contains data collected from 46 sensors deployed in the Intel Berkeley

[4] in our experiments we used the binary search.

Research Lab [27]. Mica2Dot sensors with weather boards collect timestamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. The data was collected using TinyDB in-network query processing system built on the TinyOS platform. In our experiments, we used a file that includes a log of about 2.3 million readings collected from these sensors. For the sake of simplicity, in this paper we are interested in one field of sensor measurements: the temperature. We assume that all nodes send their data to a common CH placed at the center of the Lab. First, each node reads periodically real measures while applying the member node aggregation. At the end of this step, each node sends its set of measures/weights to the CH which in turn applies CH aggregation to theses sets. Furthermore, we compare our technique to the PFF technique proposed by Bahi, J. et al. [21] with two values of the Jaccard similarity threshold t (0.75 and 0.8). We have implemented both techniques on a java simulator and we compared the results of 15 periods in all the experiments.

We evaluated the performance using the following parameters:

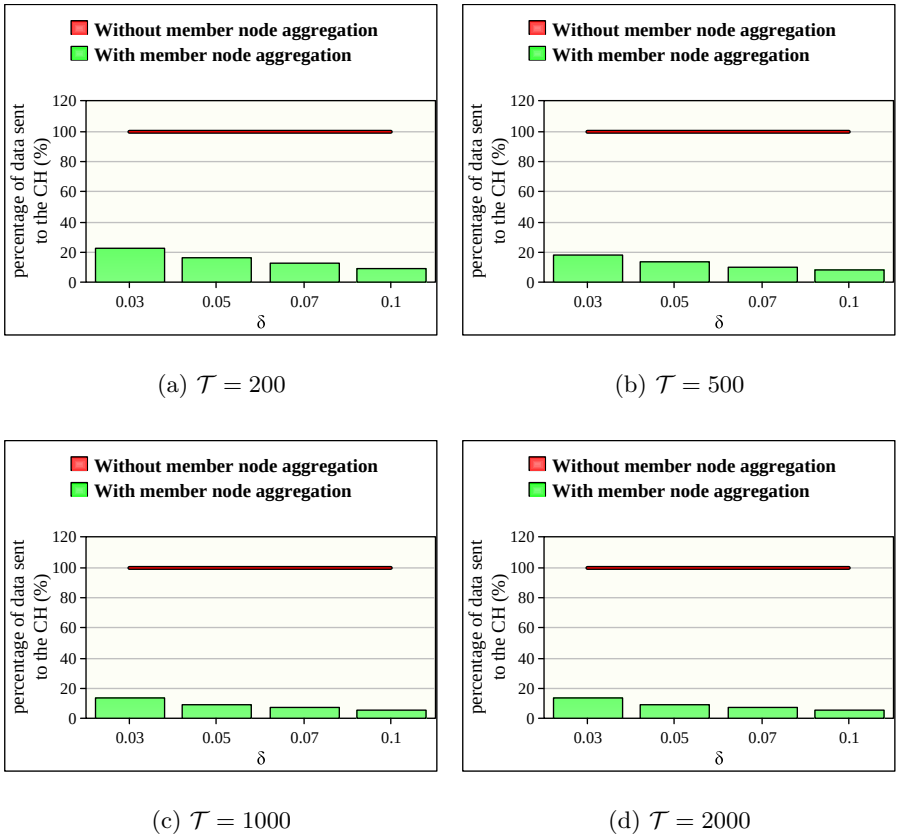
- δ, which defines the *Similar* function between two measurements. We varied δ to : 0.03, 0.05, 0.07 and 0.1.
- τ, the number of sensor measurements taken by each sensor node during a period. We varied τ to: 200, 500, 1000 and 2000.
- α, the false-rejection probability in the Anova model which we varied to 0.01 and 0.05.

5.1 Percentage of data sent to the CH

In the first aggregation level, each member node searches the similarity between measures captured at each period, using the *Similar* function, and assigns for each measure its weight. Therefore, the result of the aggregation in this level depends on the chosen threshold δ, and the number of the collected measures in period τ. Fig. 3 shows the percentage of data sent by each node to the CH at each period with and without applying the first aggregation level. The obtained results show that, at each period, each node reduces more than 68% the amount of collected data after the first aggregation level while it sends all the collected data, e.g. 100%, without applying this aggregation level. Therefore, our technique can successfully eliminate redundant measures at each period and reduce the amount of data sent to the CH. We can observe also that at the first aggregation level, data redundancy increases when τ or δ increases. This is because, *Similar* function will find more similar measures to be eliminated at each period.

5.2 Number of pairs of redundant sets generated at the CH

When receiving all the sets from its member nodes at the end of each period, CH applies the second aggregation level to find pairs of redundant sets. Fig. 4 shows the obtained number of pairs of redundant sets when applying one way Anova model with the three tests presented above, compared to the number of similar sets obtained when applying PFF. In Fig. 4(a and b), we fixed α to 0.01 and we



(a) $\mathcal{T} = 200$ (b) $\mathcal{T} = 500$

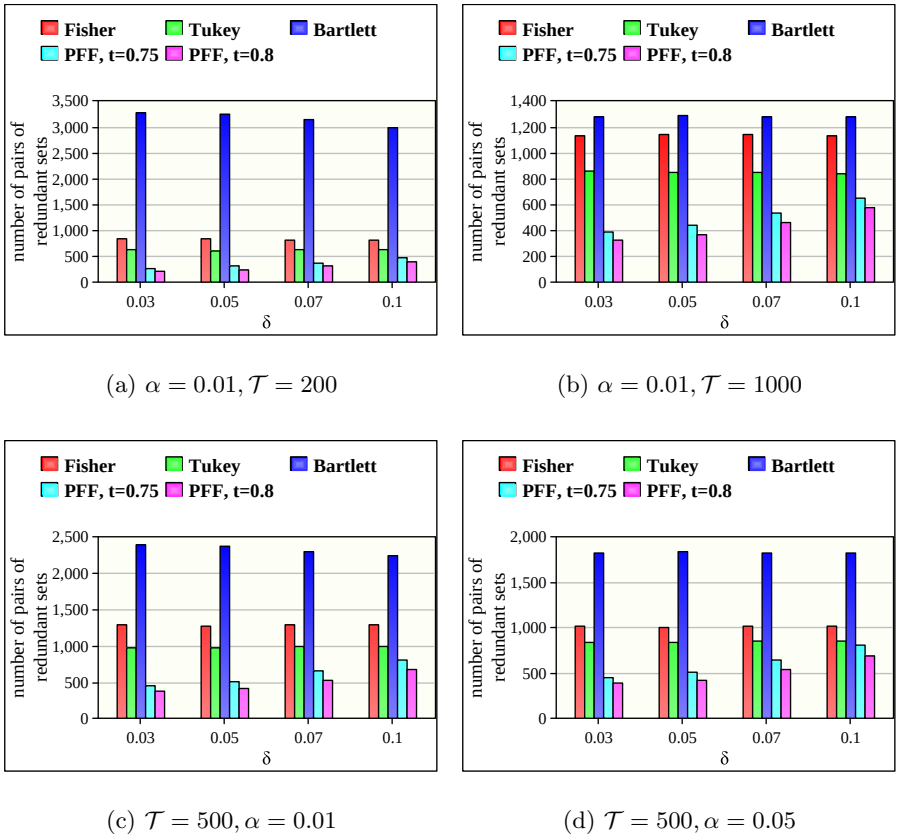
(c) $\mathcal{T} = 1000$ (d) $\mathcal{T} = 2000$

Fig. 3. Percentage of data sent to the CH.

varied $\mathcal{T}$ to 200 and 1000 respectively, while in Fig. 4(c and d) we fixed $\mathcal{T}$ to 500 and we varied α to 0.01 and 0.05 respectively. The obtained results show that, CH finds more redundant sets when applying our technique in all the cases. This is because, the variance condition in the one way Anova model is more flexible compared to the similarity condition used in PFF.

Based on the obtained results, we can also deduce:

- Bartlett test finds more pairs of redundant sets compared to Tukey and Fisher tests. This is because Bartlett test is more flexible regarding the variance between measures (Equation (5)) compared to the variance calculated in Fisher (Equation (2)) and Tukey (Equations (3) and (4)).
- The obtained number of pairs of redundant sets decreases in the three tests when α increases. This is because, when the risk α increases the null hypothesis will have higher probability of being rejected.



(a) $\alpha = 0.01, \mathcal{T} = 200$ (b) $\alpha = 0.01, \mathcal{T} = 1000$

(c) $\mathcal{T} = 500, \alpha = 0.01$ (d) $\mathcal{T} = 500, \alpha = 0.05$

Fig. 4. Number of pairs of redundant sets.

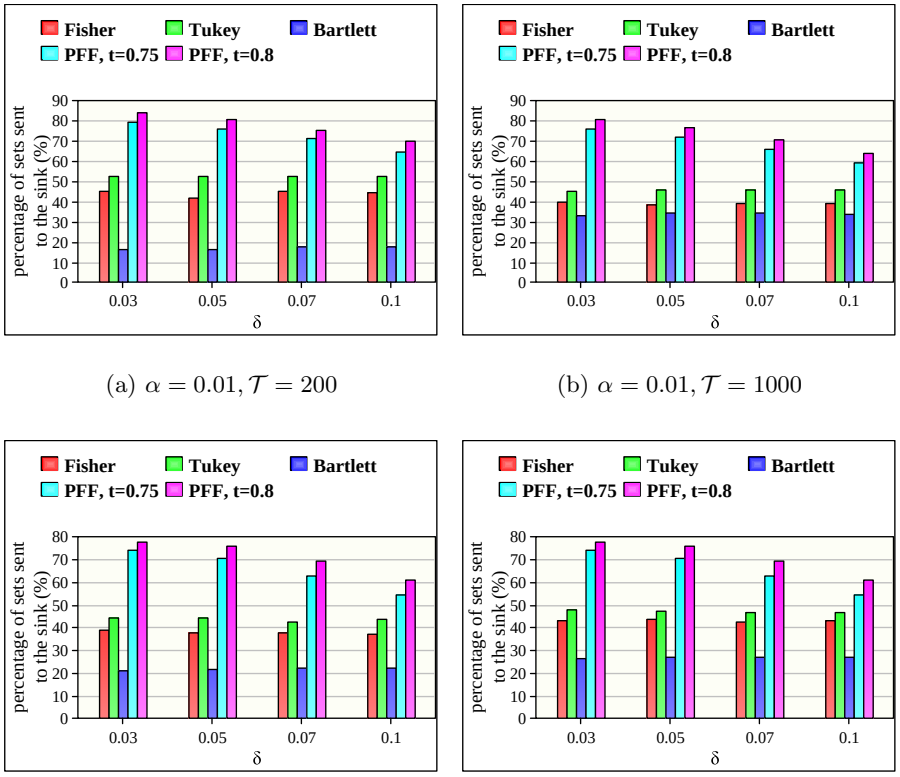
5.3 Percentage of sets sent to the sink

In this section, our objective is to show how the CH is able to eliminate redundant sets at each period using redundant data reduction algorithm. Fig. 5 shows the percentage of the remained sets that will be sent to the sink after eliminating the redundancy. Fig. 5(a and b) show the results when we fixed α to 0.01 and varied $\mathcal{T}$ to 200 and 1000 respectively, while Fig. 5(c and d) show the results when we varied α to 0.01 and 0.05 and fixed $\mathcal{T}$ to 500. We can show clearly that, our technique sends much less sets at each period to the sink with the different parameters. This is because, CH found more redundant sets using the variance condition (Fig. 4).

Based on the obtained results, we can also deduce:

- Bartlett test sends the less percentage of sets to the sink since it found more redundant sets compared to Fisher and Tukey tests (see Fig. 4).

- The percentage of sets sent to the sink for the three tests is almost fix when fixing $\mathcal{T}$ and increasing δ. This is because, the data set saves the same variance when changing δ.
- CH eliminates more redundant sets in the three tests when decreasing α. This is because when α decreases, the number of pairs of redundant sets increases (see Fig. 4(c and d)).



(a) $\alpha = 0.01, \mathcal{T} = 200$

(b) $\alpha = 0.01, \mathcal{T} = 1000$

(c) $\mathcal{T} = 500, \alpha = 0.01$

(d) $\mathcal{T} = 500, \alpha = 0.05$

Fig. 5. Percentage of sets sent to the sink.

5.4 Data accuracy

Eliminating redundant data without losing accuracy is an important challenge for the WSN. Data accuracy represents the measure "loss rate" taken by sensor nodes and not received by the sink [21]. Since CH merges the weights of similar measures in the redundant sets in one record compared to PFF which removes one between them, the integrity of the information is totally saved in

our technique. This fact is obtained independently from the values of τ, δ and α, whereas the percentage of loss measures in PFF can up to 5.4 for some values of the parameters [21]. Therefore, we can consider that our technique decreases the amount of redundant data forwarded to the sink without any loss of information integrity.

5.5 Energy consumption at the CH

In this section, our objective is to study the energy cost at the CH level. Therefore, we used the same radio model as discussed in [27]. In this model, a radio dissipates $E_{elec} = 50\,nJ/bit$ to run the transmitter or receiver circuitry and $\beta_{amp} = 100\,pJ/bit/m^2$ for the transmitter amplifier. Radios have power control and can expend the minimum required energy to reach the intended recipients as well as they can be turned off to avoid receiving unintended transmissions. Equations used to calculate transmission costs and receiving costs for a k-bit messages and a distance d are respectively shown in Equations (7) and (8):

$$E_{TX}(k,d) = E_{elec} \times k + \beta_{amp} \times k \times d^2 \tag{7}$$

$$E_{RX}(k) = E_{elec} \times k \tag{8}$$

Recall that the CH will receive n data sets coming from its member nodes at each period. The size of each set is equal to the number of measures sent in addition to the number of weights sent. We consider that each measure or weight is equal to 64 bits. Therefore, the energy consumption at the second level will be equal to the energy consumed when the CH receives the data sets from its member in addition to the energy consumed when it sends them after the aggregation. Consequently, after 15 periods as we calculated in our experiments, the total energy consumption at the CH is calculated as shown in Equation (9)

$$E_{CH}(m,d) = E_{RX_{total}} + E_{TX}(m,d) = \Big(2 \times 64 \times E_{elec} \times \sum_{i=1}^{n} |M_i|\Big) + \Big(64 \times E_{elec} \times m + 64 \times \beta_{amp} \times m \times d^2\Big) \tag{9}$$

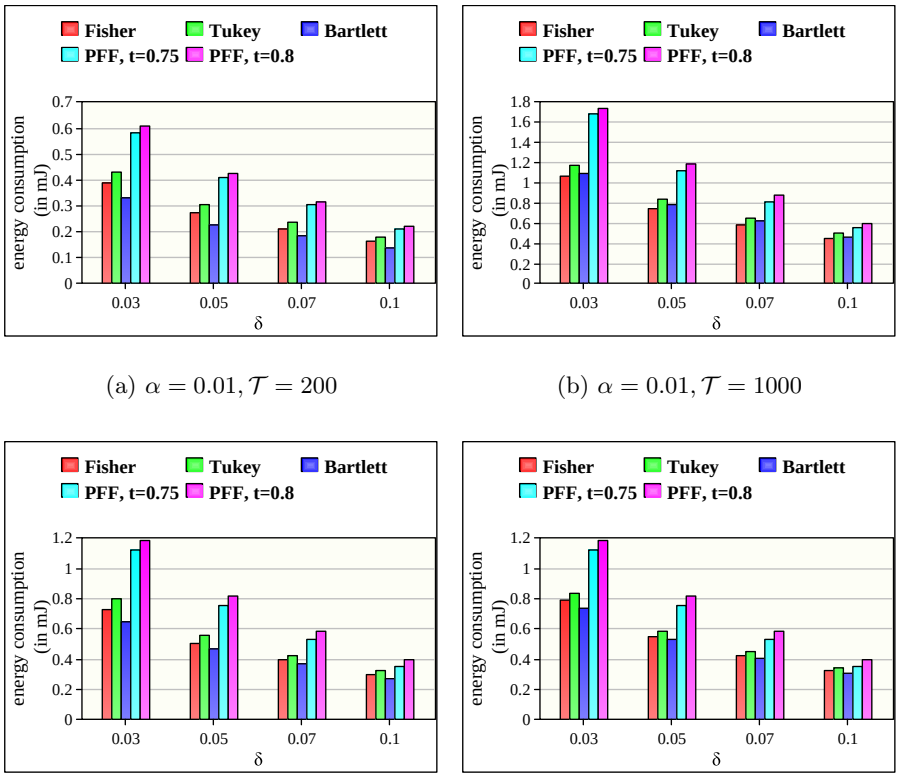
where m is the total number of the measures with their weights after the aggregation in all the sets and d is the distance between the CH and the sink.

Fig. 6 shows the energy consumption comparison between our technique and the PFF at the CH level when fixing α and varying τ (Figs. a and b) and when fixing τ and varying α (Figs. c and d). The obtained results show that our technique minimizes the energy consumption of the CH up to 45% when compared to the PFF. These results are obtained due to the fact that our technique eliminates more redundant sets compared to PFF (see Fig. 5). Therefore, we can

consider that our technique decreases the amount of redundant data forwarded to the sink and performs an overall lossless process in terms of information and integrity by conserving the weight of each measure.

Based on the obtained results, we can also deduce:

- Bartlett test decreases energy consumption of the CH more than the other tests.
- The energy consumption at the CH is more minimized when α decreases. This is because, when α decreases the percentage of sets sent to the sink decreases (see Fig. 5(c and d)).



(a) $\alpha = 0.01, \mathcal{T} = 200$ (b) $\alpha = 0.01, \mathcal{T} = 1000$

(c) $\mathcal{T} = 500, \alpha = 0.01$ (d) $\mathcal{T} = 500, \alpha = 0.05$

Fig. 6. Energy consumption at the CH.

5.6 Discussion

In this section, we discuss the results for the three tests used with ANOVA model in terms of conserving energy of the sensors. First, by fixing α and varying $\mathcal{T}$ as

shown in Fig. 6(a, b and c), we can deduce that Bartlett test allows more energy saving than Fisher and Tukey tests when the period is small (e.g. $\mathcal{T}$ equals to 200 and 500 in Figs. a and c). Contrarily, Fisher test gives better results for large periods, e.g. $\mathcal{T}$ is greater than 1000 in Fig. c. This is because, Bartlett test is more flexible regarding the variance between measures in small periods (Equation (5)) while Fisher test is more flexible in large periods (Equation (2)).

On the other hand, by fixing $\mathcal{T}$ and varying α as shown in Fig. 6(c and d), the energy consumption is more minimized in the three tests when α is small, e.g. $\alpha = 0.01$ in Fig. c. This is because, the energy consumption highly depends on the number of pairs of redundant sets eliminated which increases when α decreases. Consequently, the null hypothesis will have higher probability of being rejected when α decreases. Furthermore, the general trend observed that is Bartlett test gives better results, in terms of energy consumption, when $\mathcal{T}$ is small while Fisher test gives better results when $\mathcal{T}$ is large.

6 Conclusion and Future Work

In this paper, we proposed a new technique for data aggregation in PSN that enforces both energy consumption and integrity of the aggregated data. Our proposed technique consists of two-level of data aggregation which applies at each cluster in a clustering network architecture. The first level is applied at the node itself to eliminate redundancy from the collected raw data before sending them to the CH. At the second level, CH searches nodes that generate redundant data sets based on the dependence of conditional variance with three different Anova tests. Comparing to other existing data aggregation techniques, experimental results on real sensor data show the effectiveness of our technique in terms of energy consumption and information integrity.

A direction for future work is to adapt our proposed technique to take into consideration reactive periodic sensor networks, where sensor nodes operate with different sampling rate. In periodic applications the dynamics of the monitored condition or process can slow down or speed up; and to save more energy the sensor node can adapt its sampling rates to the changing dynamics of the condition or process.

References

1. Abbasi, A. and Younis, M.: A survey on clustering algorithms for wireless sensor networks, Journal of Computer Communications, Vol. 30, Iss. 14-15, pp. 2826-2841, 2007.
2. Rozyyev, A., Hasbullah, H. and Subhan, F.: Indoor child tracking in wireless sensor network using fuzzy logic technique, Research Journal of Information Technology, Vol. 3, Iss. 2, pp. 81-92, 2011.
3. Sabri, N., Aljunid, S.A., Ahmad, R.B., Yahya, A., Kamaruddin, R. and Salim, M.S.: Wireless sensor actor network based on fuzzy inference system for greenhouse climate control, Journal of Applied Sciences, Vol. 11, Iss. 17, pp. 3104-3116, 2011.

4. Aslan, Y.E., Korpeoglu, I. and Ulusoy, O.: A framework for use of wireless sensor networks in forest fire detection and monitoring, Comput. Environ. Urban Syst., Vol. 36, Iss. 6, pp. 614-625, 2012.
5. Qian, H., Sun, P. and Rong, Y.: Design Proposal of Self-Powered WSN Node for Battle Field Surveillance, Energy Proced., Vol. 16, Part B, pp. 753-757, 2012.
6. Padmavathi, G., Shanmugapriya, D. and Kalaivani, M.: A Study on Vehicle Detection and Tracking Using Wireless Sensor Networks. Wirel. Sens. Netw., Vol. 2, No.2, pp. 173-185, 2010.
7. Di Pietro, R., Michiardi, P. and Molva, R.: Condentiality and integrity for data aggregation in WSN using peer monitoring, Security Comm. Networks, Vol. 2, Iss. 2, pp. 181-194, 2009.
8. Azhar, M., Ke, S., Shaheen, K. and Mi, X.: Data Mining Techniques for Wireless Sensor Networks: A Survey, International Journal of Distributed Sensor Networks, Vol. 2013, Iss. 2013, 24 pages, 2013.
9. Mirhadi, P., Zandinia, S., Goodarzipour, A., Salimi, S. and Goodarzipour, H.: IP2P K-means: an efficient method for data clustering on sensor networks, Management Science Letters, Vol. 3, Iss. 3, pp. 967-972 , 2013.
10. Yuan, F., Zhan, Y. and Wang, Y.: Data Density Correlation Degree Clustering Method for Data Aggregation in WSN, Sensors Journal, IEEE, Vol. 14, Iss. 4, pp. 1089-1098, 2014.
11. Enam, R.N., Qureshi, R. and Misbahuddin, S.: A Uniform Clustering Mechanism for Wireless Sensor Networks, International Journal of Distributed Sensor Networks, Vol. 2014, Iss. 2014, 14 pages, 2014.
12. Tripathi, A., Gupta, S. and Chourasiya, B.: Survey on Data Aggregation Techniques for Wireless Sensor Networks, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Iss. 7, 2014.
13. Nokhanji, N. and Hanapi, Z.M.: A Survey on Cluster-based Routing Protocols in Wireless Sensor Networks, Journal of Applied Sciences, Vol. 14, Iss. 18, pp. 2011-2022, 2014.
14. Zou, P. and Liu, Y.: A Data-aggregation Scheme for WSN based on Optimal Weight Allocation, Journal Of networks, Vol. 9, No. 1, pp. 100-107, 2014.
15. Tran, K.T-M. and Oh, S.-H.: A Data Aggregation Based Efficient Clustering Scheme in Underwater Wireless Sensor Networks, Ubiquitous Information Technologies and Applications, Lecture Notes in Electrical Engineering, Vol. 280, pp 541-548, 2014.
16. Kumar, S., Prateek, M., Ahuja, N.J. and Bhushan, B.: MEECDA: Multihop Energy Efficient Clustering and Data Aggregation Protocol for HWSN, International Journal of Computer Applications, Vol. 88, No. 9, pp. 28-35, 2014.
17. Chao, C.-M. and Hsiao, T.-Y.: Design of structure-free and energy-balanced data aggregation in wireless sensor networks, Journal of Network and Computer Applications, Vol. 37, pp. 229-239, 2014.
18. Li, G. and Wang, Y.: Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks, EURASIP Journal on Wireless Communications and Networking, Vol. 2013, Iss. 85, 2013.
19. Shan, M., Chen, G., Luo, D., Zhu, X., and Wu, X.: Building Maximum Lifetime Shortest Path Data Aggregation Trees in Wireless Sensor Networks, Journal ACM Transactions on Sensor Networks (TOSN), Vol. 11, Iss. 1, Article 11, 2014.
20. Shim, Y. and Kim, Y.: Data Aggregation with multiple sinks in Information-Centric Wireless Sensor Network, International Conference on Information Networking (ICOIN 2014), pp. 13-17, 2014.

21. Bahi, J., Makhoul, A. and Medlej, M.: A Two Tiers Data Aggregation Scheme for Periodic Sensor Networks, Ad Hoc & Sensor Wireless Networks, Vol. 21, Iss. (1-2), pp. 77-100, 2014.
22. Harb, H., Makhoul, A., Tawil, R. and Jaber, A.: A Suffix-Based Enhanced Technique for Data Aggregation in Periodic Sensor Networks, 10th IEEE Int. Wireless Communications and Mobile Computing Conference (IWCMC 2014), pp 494-499, 2014.
23. Harb, H., Makhoul, A., Laiymani, D., Jaber, A. and Tawil, R.: K-Means Based Clustering Approach for Data Aggregation in Periodic Sensor Networks, 10th IEEE Int. Conf. on Wireless and Mobile Computing, Networking and Communications (WIMOB 2014), pp. 434-441, 2014.
24. Laiymani, D. and Makhoul, A.: Adaptive data collection approach for periodic sensor networks, 9th International Wireless Communications and Mobile Computing Conference (IWCMC), pp.1448-1453, 2013.
25. Hall, R.: http://web.mst.edu/ psyworld/tukeysexample.htm, Psychology World, 1998.
26. Snedecor, G. and Cochran, G.: Statistical Methods, Eighth Edition, Iowa State University Press, 1989.
27. Samuel Madden. http://db.csail.mit.edu/labdata/labdata.html.