An Enhanced K-means and ANOVA-based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks

Hassan Harb, Abdallah Makhoul, and Raphaël Couturier,

***Abstract*—Underwater wireless sensor networks (UWSNs) have recently been proposed as a way to observe and explore aquatic environments. Sensors in such networks are used to perform pollution monitoring, disaster prevention or assisted navigation and to send monitored data to the sink. Compared to traditional sensor networks, sensors in UWSNs consume more energy due to the acoustic technology used in under water communications. Node clustering is a common method to organize data traffic and reduce in-network communications while improving scalability and energy consumption. In this paper, we present a new clustering method to handle the spatial similarity between node readings. We suppose that readings are sent periodically from sensor nodes to their appropriate cluster-heads (CHs). Then a two tier data aggregation technique is proposed. At the first level, each node periodically cleans its readings in order to eliminate redundancies before sending its data set to its CH. Once the CH receives all data sets, it applies an enhanced K-means algorithm based on a one-way ANOVA model to identify nodes generating identical data sets and to aggregate these sets before sending them to the sink. Our proposed approach is validated via experiments on real sensor data and comparison with other existing clustering and data aggregation techniques.**

***Index Terms*—Underwater Wireless Sensor Network (UWSN), data aggregation, one-way ANOVA model, hierarchical k-means clustering.**

I. Introduction

UNDERWATER wireless sensor network (UWSN) is a special kind of wireless sensor network, which is composed of underwater acoustic sensor nodes. UWSNs are deployed in an underwater or aquatic environment and are capable of monitoring nearby surroundings. They represent the solution for many different applications such as real-time warship monitoring, locating mooring positions and submerged wrecks, oceanographic data collection, disaster prevention, etc [1], [2]. Underwater sensor nodes are small devices with constrained energy and little memory [3]. Moreover, these nodes use acoustic signals that can travel to longer distance than radio waves due to lower frequency [4]–[6]. Hence, unlike traditional sensor networks, sensors in UWSNs consume more energy due to the acoustic technology used in under water communications. In addition, they are costly and difficult to replace. Therefore, there are increasing demands for innovative methods to improve energy efficiency and to prolong the network lifetime. The node clustering and the data aggregation at the level of cluster heads (CHs) are two common methods to organize data traffic and reduce in-network redundancies while improving scalability and energy consumption. Indeed, nodes clustering makes a network look smaller and extends its lifetime by reducing data transmissions between the nodes and the sink [7]; while data aggregation is considered to be the best way to minimize the energy consumption by eliminating redundant data received from sensor nodes, and to reduce the number of transmissions to the sink. Thus, combining the two techniques will lead to the enhancement of the network performances.

Subsequently, many clustering protocols have been proposed in UASN [6]. However, few protocols consider sensor readings similarity and the correclation between received data. Thus, the lack of suitable energy efficient protocols for handling such correlations leads us to study a data aggregation and clustering protocol that creates clusters of nodes with identical readings. In underwater applications the sensor nodes are usually deployed over large areas with the purpose of periodically sensing nearby surroundings and transmitting data to a sink or base station. In our approach, we consider that each cluster is formed by a cluster head and several member nodes. The role of a cluster head is to collect data from its node members, aggregating these data before sending them to the sink. Therefore, sensor nodes send their readings to the CH that performs data aggregation in a periodic manner. In this paper, we propose a two-tier data aggregation technique to preserve energy in a UWSN. On the one hand, the first tier is at the sensor node level, where each node is responsible for cleaning and eliminating redundancies from the data collected by the node at each period. We provide a simple algorithm based on a distance function called $link$ between sensor's readings. On the other hand, the second tier is applied at the CH level where a data aggregation and clustering technique is proposed. After receiving all data sets from the member nodes, each CH applies a k-means based clustering method to classify these sets by identical data sets with the aim to eliminate redundancies and reduce the huge amount of data. For this purpose, we propose an enhanced k-means clustering method using the one-way ANOVA model and three different statistical tests in order to identify neighboring nodes generating identical data sets. Finally, we study the performance of our proposed technique while using real underwater sensor readings. We show via simulations the effectiveness of using

H. Harb is with FEMTO-ST Laboratory, the DISC department, University of Franche-Comté, Belfort, France, and with the Department of Computer Science, Lebanese University, Beirut, Lebanon, e-mail: hassan.moustafa_harb@univ-fcomte.fr.

A. Makhoul and R. Couturier are with FEMTO-ST Laboratory, the DISC department, University of Franche-Comté, Belfort, France, e-mail: firstname.lastname@univ-fcomte.fr.

data similarity and the analysis of variance to reduce the packets size, to minimize data redundancy, and to decrease the energy consumption of the network.

The rest of the paper is organized as follows: related works are discussed in Section II. Section III describes the cluster-based architecture used in our technique, the scenario and some definitions.The ANOVA model and statistical tests are described in Section IV. Section V presents the aggregation phase at the CH, based on K-means algorithm adapted to ANOVA model. Experimental results are presented in Section VI. Finally, Section VII concludes our paper and gives some perspectives.

II. Related Work

In UASN, a lot of researches have been proposed for data aggregation based on clustering scheme aiming at minimizing energy consumption and extending the network lifetime. The idea behind this approach is to avoid multihop communications and to build an aggregate path over the network. The authors in [12] present a review of various data aggregation techniques and clustering schemes proposed recently by different researches in underwater sensor networks. In [13], the authors propose a data aggregation technique based on clustering which involves four phases. The main goal of these phases includes reducing the energy consumed by the overall network, increasing the throughput, and minimizing data redundancy while still guarantying data accuracy. The authors in [14] propose to design a fuzzy based clustering and aggregation technique for UWSN. In this technique the parameters residual energy, distance to sink, node density, load and link quality are considered as input to the fuzzy logic. Based on the output of fuzzy logic module, appropriate cluster heads are elected and act as aggregator nodes. The authors in [15] propose EBDSC, a distributed Energy-Balanced Dominating Set-based Clustering scheme, to prolong the network lifetime by balancing energy consumption among different nodes. In EBDSC, a node becomes a candidate cluster head if it has the longest lifetime among its neighbors. In [16], the authors develop an architecture which can be used to build different networks with different routing protocols. One of the main advantages of that architecture is that if all CHs switch off at the same time, the system is able to continue working. In [17], the authors propose a Cluster-based False data Filtering Scheme (CFFS) that can detect and filter out false reports travel in the network before leading to a waste of energy of this network. In [18], the authors try to capitalize the delay-tolerance of various applications with the aim of reducing the energy consumption. The proposed approach relies on Adaptive Modulation and Coding (AMC) to dynamically change the modulation and coding scheme. In [19], the authors propose RTOC (Real-Time Opportunistic Coding), a new XOR-based opportunistic network coding architecture for real-time data transmission. RTOC is an application-independent architecture that takes into consideration the characteristics of real-time traffic and provides an efficient framework to optimize multimedia application requirements such as bandwidth, delay and loss.

In other studies, [20]–[22], the authors suggest using similarity functions for data aggregation in cluster-based periodic sensor networks. The main objective is to eliminate redundancy and reduce the size of data transmitted in order to optimize the energy consumption and to reduce overload on the network level. The first level at sensor nodes, called local aggregation, with which each sensor node sends, at each period p, its aggregated set of data to the aggregator. At the second level, a prefix frequency filtering (PFF) technique is provided to identify all pairs of neighbor nodes generating similar sets of data [20]. Then several optimizations of the PFF technique [21] have been proposed in order to avoid comparing all received sets thus minimizing data latency. In [22] the authors use Euclidean distance and cosine distance at the aggregator level to build an efficient underwater network by reducing packet size and by minimizing data redundancy. Although all similarity functions allow eliminating redundancy among data, it remains a difficult technique for the aggregator in terms of data latency, due to the comparison of each pair of sets, and the energy consumption. In this paper, we propose a data aggregation and clustering technique in order to eliminate further redundancies, to enhance data latency and to optimize the energy consumption of the whole network.

III. Cluster-based Architecture for UASN

In this paper, we consider a 2-D UASN with cluster-based architecture for the network. As mentioned before, clustering is considered to be an efficient topology control method which can increase network scalability and lifetime. In clustering schemes, the network is divided into a number of clusters based on certain rules where each cluster has a Cluster-Head (CH). CH is responsible for managing the cluster. Data transmission between sensor nodes and their appropriate CHs is based on a single-hop communication. In our work, we consider the periodic data collection model, where each sensor node periodically sends (within a period p) its data to the appropriate CH which, in turn, sends it to the sink (Fig. 1). Then, we propose a two-tier data aggregation technique aiming at reducing the redundancy among data transmitted over the network thus extending network lifetime. The sensor nodes form the first tier while the CH represents the second one. Our proposed technique efficiently reduces the amount of data sent to the sink while a minimum percentage of received data is guaranteed.

A. Scenario and definitions

In PUASN, a period p_j is divided into time slots $s_{j\mathcal{T}}$ where $\mathcal{T}$ is the total number of slots in the period p_j. Each sensor node n takes a new measure m_{ji} at each slot s_{ji}, where $i \in [1,\mathcal{T}]$, then it forms a vector of measures during the period p_j as follows: $M_n = [m_{j1}, m_{j2}, \ldots, m_{j\mathcal{T}}]$. Fig. 2 shows an example of PUASN where each sensor node takes five measures (e.g. $\mathcal{T}$=5), at each period p_j ($j \in [1,3]$) and sends its set of collected data $M = [m_{j1}, m_{j2}, m_{j3}, m_{j4}, m_{j5}]$ to the CH at the end of the period.

Usually the collected measures are highly dependent on the monitored condition. Subsequently, the dynamic of the monitored conditions can slow down or speed up [23]. Hence, the nodes may take the same or very similar measures several

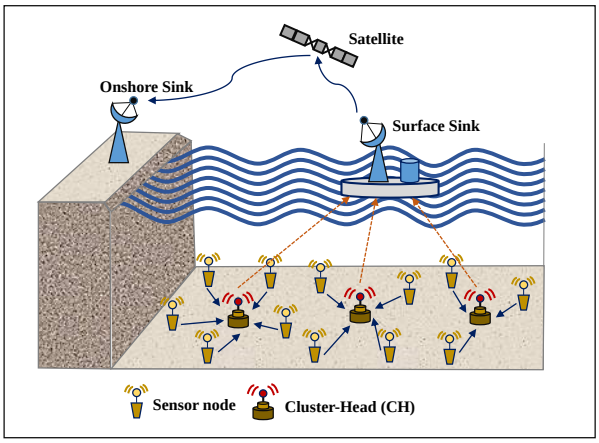


Fig. 1. Cluster-based network architecture for 2-D UASN.

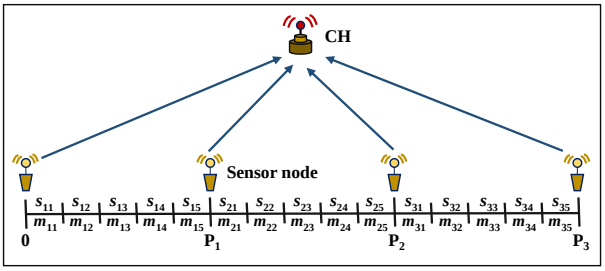


Fig. 2. Illustrative example of periodic UASN (PUASN).

times, especially when slots are short. In order to eliminate redundant values from the set M_n, the node n searches for data similarity in the set. Thus, to identify the similarities between two measures, we define the two following functions:

Definition 1 (Similar function): We define the $Similar$ function between two measurements captured by the same sensor node n as:

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

where m_i and $m_j \in M_n$ and 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, $wgt(m_i)$): The weight of a measurement m_i is defined as the frequence of the similar or the same measurements (according to the $Similar$ function) in the same set.

For each new captured measurement (at each time slot), a sensor node n searches for similarities of the new taken measurement. If a similar measurement is found, it deletes the new one and increments the corresponding weight by 1, or else it adds the new measure to the set and initializes its weight to 1. This leading to the following definitions:

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

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

For the sake of simplicity, in this paper we consider that all sensor nodes operate at the same sampling rate and each node captures τ measures in each period[1]. Thus we can deduce that for every received set M_n from node n we have: $Card_w(M_n) = \tau$.

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

IV. Variance Study

Studying the variance between measurements in the data sets is an effective way to find nodes that generate redundant data. The ANOVA model provides a statistical tests of whether or not the means of several sets are equal. In the typical application of ANOVA, the null hypothesis (H_0) supposes that the variance between sets is not significant. Consequently, the test result (R) of the ANOVA is the ratio of the computed variance based on the measurements in the sets. R can be calculated in different manners depending on the statistic tests (presented in the next section) proposed in the ANOVA model. The sets are considered duplicated if the result R is less than a threshold T (significance level) for some desired false-rejection probability (risk α).

At each period, we suppose that the CH receives n sets from its sensor nodes, each set contains τ measures. Also, we assume that measures in each set M_j are independent, with mean $\overline{Y_j}$ and that the variances of sets are equal $\sigma_n^2 = \sigma^2$. Then the measure's variables can be written as follows:

$$m_{ji} = \overline{Y_j} + \epsilon_{ji}; \qquad j = 1, \ldots, n; \qquad i = 1, \ldots, |M_j|$$

Where ϵ_{ji} are the residuals which are independent and are normally distributed following $N(0, \sigma^2)$.

For each set M_j, we denote by $\overline{Y_j}$ its mean, σ_j^2 its variance and $\overline{Y}$ the mean of all the n sets respectively as follows:

$$\overline{Y_j} = \frac{1}{Card_w(M_j)} \sum_{k=1}^{|M_j|} \big(m_{jk} \times wgt(m_{jk})\big),$$

$$\sigma_j^2 = \frac{1}{Card_w(M_j)} \sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times (m_{jk} - \overline{Y_j})^2\big),$$

$$\overline{Y} = \sum_{j=1}^{n} \sum_{k=1}^{|M_j|} \Big(\frac{1}{Card_w(M_j)} \times \big(m_{jk} \times wgt(m_{jk})\big)\Big)$$

where $m_{jk} \in M_j$.

Since $Card_w(M_1) = \ldots = Card_w(M_j) = \ldots = Card_w(M_n) = \tau$:

$$\overline{Y_j} = \frac{1}{\tau} \sum_{k=1}^{|M_j|} \big(m_{jk} \times wgt(m_{jk})\big),$$

[1]Note that it is possible to take different τ for sensors.

$$\sigma_j^2 = \frac{1}{\mathcal{T}} \sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times (m_{jk} - \overline{Y_j})^2\big),$$

$$\overline{Y} = \frac{1}{\mathcal{T}} \sum_{j=1}^{n} \sum_{k=1}^{|M_j|} \Big(m_{jk} \times wgt(m_{jk})\Big)$$

where $m_{jk} \in M_j$.

The total variation is the sum of the variation (SR) within each set and the variation (SF) between the sets. The basic idea is to calculate the mean of the measurements within each set, then to compare the variance among these means with the average variance within each set. Under the null hypothesis that the measurements in the different sets all have the same mean, the weighted among-sets variance will be the same as the within-sets variance. As the means get further apart, the variance among the means increases. The statistical test is thus the ratio of the variance among means divided by the average variance within sets, or F_s. This statistical test has a known distribution under the null hypothesis, so the probability of obtaining the observed F_s under the null hypothesis can be calculated.

A. Statistical Tests

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

1) Fisher Test/F-test: The result of the F-test is calculated by the following formula:

$$R = \frac{SF/(n-1)}{SR/(n \times (\mathcal{T}-1))} = \frac{\big(\mathcal{T} \times \sum_{j=1}^{n} (\overline{Y_j} - \overline{Y})^2\big)/(n-1)}{\big(\sum_{j=1}^{n} \sum_{k=1}^{|M_j|} (wgt(m_{jk}) \times (m_{jk} - \overline{Y_j})^2)\big)/(n \times (\mathcal{T}-1))} \quad (1)$$

At the end of each period, the CH receives n sets from its sensor nodes and it will test the hypothesis that all the means of sets are the same or not. If the hypothesis is correct then, R will have a Fisher distribution, with $F(n-1, n \times (\mathcal{T}-1))$ degrees of freedom. The hypothesis is rejected if the R calculated from the measures is greater than the critical value of the F distribution for some desired false-rejection probability (risk α). Let $T = F_{1-\alpha}(n-1, n \times (\mathcal{T}-1))$. The decision is based on R and T :

- if $R > T$ then the hypothesis is rejected with false-rejection probability α, and the variance between sets are significant.
- if $R \leq T$ the hypothesis is accepted.

2) Tukey Test:

Tukey's test [24] can be applied to n data sets based on the following equations:

$$SS_{total} = \sum_{j=1}^{n} \sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times m_{jk}^2\big) - \frac{\Big(\sum_{j=1}^{n} \sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times m_{jk}\big)\Big)^2}{n \times \mathcal{T}} \quad (2)$$

$$SS_{among} = \frac{\sum_{j=1}^{n} \Big(\sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times m_{jk}\big)^2\Big)}{\mathcal{T}} - \frac{\Big(\sum_{j=1}^{n} \sum_{k=1}^{|M_j|} \big(wgt(m_{jk}) \times m_{jk}\big)\Big)^2}{n \times \mathcal{T}} \quad (3)$$

$$SS_{within} = SS_{total} - SS_{among};\ df_{among} = n-1;$$
$$df_{within} = n \times (\mathcal{T}-1);\ MS_{among} = \frac{SS_{among}}{df_{among}};$$
$$MS_{within} = \frac{SS_{within}}{df_{within}};\ R = \frac{MS_{among}}{MS_{within}}$$

Where:

- n : Number of total sets,
- $\mathcal{T}$: Number of measures in each set,
- SS_{within} : Sum of squares within the n sets,
- SS_{among} : Sum of squares between the n sets,
- MS_{within} : Mean squares within the n sets,
- SS_{among} : Sum of squares between the n sets.

Therefore, when we calculate the result of the Tukey, e.g. R, we check to see if R is statistically significant on the probability table with appropriate degrees of freedom $T = df(df_{among}, df_{within})$. The decision is based on R and T:

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

3) Bartlett Test: To investigate the significance of the differences between the variances of n normally distributed sets, let σ_j^2 denote the variance of the set M_j where $j = 1, ..., n$. Bartlett Test [25] has the following expression:

$$R = \frac{(\mathcal{T}-1)(n \times \ln(\sigma_p^2) - \sum_{j=1}^{n} \ln(\sigma_j^2))}{\lambda} \quad (4)$$

where :

$$\lambda = 1 + \frac{(n+1)}{3 \times n \times (\mathcal{T}-1)} \quad (5)$$

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

$$\sigma_p^2 = \frac{1}{n \times (\mathcal{T}-1)} \times \sum_{j=1}^{n} \sigma_j^2$$

Bartlett's test has a $(n-1)$ degrees of freedom. Thus the null hypothesis is rejected if $R > T_{n-1,\alpha}$ (where $T_{n-1,\alpha}$ is the upper tail critical value for the T_{n-1} distribution). We suppose that $T = T_{n-1,\alpha}$, thus the decision is based on the following rule:

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

B. Variance Study based Algorithm

In our technique, the one-way ANOVA model is used to identify if a set of data sets has a low variance between their measures or not (Algorithm 1). The algorithm uses the set of data sets which will be tested and returns a boolean value indicating that the sets are redundant or not. First, it calculates the corresponding R result as described in each test presented before. Then, it searches the corresponding threshold T based on the probability table for each test with the appropriate degrees of freedom (line 2). Finally, it concludes that the data sets are redundant only if the variance between their measures (R) is less than the threshold T (lines 3 and 4).

Algorithm 1 Variance Study Algorithm.

Require: Set of measures' sets $M = \{M_1, M_2 ... M_n\}$.
Ensure: Boolean value: $true$ or $false$.
1: compute R for M
2: find T
3: **if** $R \leq T$ **then**
4: return $true$
5: **else**
6: return $false$
7: **end if**

V. Aggregation at the CH

The second phase of the aggregation is done at the CH level where at the end of each period it receives the sets of measurements with their weights from all its node members. The main objective at this level is to identify neighboring nodes that generate the same or very similar data sets in order to reduce the amount of data to send to the sink. Hence, depending on the changes of the monitored condition, neighboring sensor nodes may collect duplicated data if they are geographically too close or send the sensed data to the CH over a short period of time. In this phase the CH uses an updated k-means algorithm based on the data variance of the received sets. It uses a one-way ANOVA model to determine duplicated sets based on the variance study. In this section, we describe our method based on this study and the k-means for data aggregation.

A. K-means Clustering Algorithm

In this section we introduce a k-means based method for clustering in UWSN. Clustering has been proved as an effective way to find sensor nodes that generate redundant data sets [26]. It allows data reduction and provides more accurate field testing information and system status information by sending only the useful information. The useful information will be used at the sink to provide correct judgment and to acquire more reasonable results. In addition, sending data will consume more energy than its own consumption.

The main objective of our clustering method is to create groups of data sets, or clusters, in such a way that data sets in the same cluster are very similar and data sets in different clusters are quite distinct. Among clustering algorithms, the K-means clustering algorithm, proposed by MacQueen three decades ago [27], is one of the best-known and most popular clustering algorithms used in a variety of domains such as scientific field research and industrial applications [28], [29].

It is relatively simple and is mainly based on the Euclidian distances to form the clusters. Generally, procedure of K-means algorithm starts with initial K cluster centroids, then it assigns each data sets to the nearest centroid, updates the cluster centroids, and repeats the process until the criterion function converges (Algorithm 2). Commonly, the K-means algorithm criterion function adopts square error criterion (E) [30], where E is the total square error of all the data sets in the cluster. The criterion function is to make the generated cluster as compacted and independent as possible.

Algorithm 2 K-means Algorithm.

Require: Set of measures' sets $M = \{M_1, M_2 ... M_n\}$, K.
Ensure: Set of clusters $C = \{C_1, C_2 ... C_K\}$.
1: **for** $i \leftarrow 1$ to K **do**
2: $C_i \leftarrow \emptyset$
3: randomly choose centroid r_i among M_j belongs to C_i
4: **end for**
5: **repeat**
6: **for** each set $M_j \in M$ **do**
7: Assign M_j to the cluster C_i with nearest r_i (i.e., $d(M_j, r_i) \leq d(M_j, r_{i*}); i \in \{1, \dots K\}$)
8: **end for**
9: **for** each cluster C_i, where $i \in \{1, \dots K\}$ **do**
10: Update the centroid r_i to be the centroid of all sets currently in C_i, so that $r_i = \frac{1}{|C_i|} \sum_{j \in C_i} d(M_j, r_i)$
11: **end for**
12: **until** all K clusters meet the criterion function convergence
13: return C

The initial number of clusters (K) is the main challenge of K-means. Therefore, several attempts such as [30]–[32] have been made to solve the cluster initialization problem by proposing either static or dynamic initialization. In the first one, the K-means algorithm assumes that the number of clusters K is already known by the users, which is not true in practice. In the second one, the K-means initializes the number of clusters to the total number of data sets then it joins the clusters that are close, according to the criterion function, to obtain at the end the optimal value of K. To the best of our knowledge, we propose in this section a new

initialization method to find dynamically the optimal number of clusters. Compared to existing dynamic methods, our proposed method assumes that all data sets are in the same cluster at the beginning then it uses the dependence between measurements in these sets based on the one-way ANOVA model to obtain the optimal number K.

1) Definitions and Assumptions: Let $N = \{N_1, N_2, \ldots, N_n\}$ denote the set of sensor nodes with their data sets $M = \{M_1, M_2, \ldots, M_n\}$ generated at each period respectively. Let $C = \{C_1, C_2, \ldots, C_K\}$ denote the final clusters where the data sets will be assigned to them, $K \leq n$. We also assume that each cluster C_i has the centroid r_i. Therefore, we provide the following two definitions that compute the Euclidean distance between the mean $\overline{Y_j}$, respectively the variance σ_j^2, of the set M_j and the centroid r_i of the cluster C_i as follows:

Definition 5 (Mean distance, $d_m(\overline{Y_j}, r_i)$): The Euclidean distance between the mean $\overline{Y_j}$ of the set M_j and the cluster centroid r_i is defined as follows:

$$d_m(\overline{Y_j}, r_i) = \sqrt{(\overline{Y_j} - r_i)^2}. \quad (6)$$

Definition 6 (Variance distance, $d_v(\sigma_j^2, r_i)$): The Euclidean distance between the variance σ_j^2 of the set M_j and the cluster centroid r_i is defined as follows:

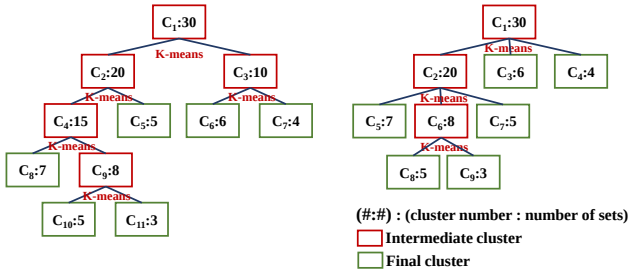
$$d_v(\sigma_j^2, r_i) = \sqrt{(\sigma_j^2 - r_i)^2}. \quad (7)$$

2) Adapted K-means Algorithm Description: When receiving all data sets from its member nodes at the end of each period, the CH considers that all the data sets are in the same cluster. Then, it begins dividing this cluster each time by applying the adapted K-means algorithm until it obtains the null hypothesis (Algorithm 3). In our approach, the criterion function of the k-means algorithm is the result of the statistical test studied before and based on the ANOVA model. Algorithm 3 describes the K-means adapted to UWSN. First, it starts, as said before, by grouping all the received sets at the initial same cluster (line 5). Then, it searches the variance between measurements in all the sets in the initial cluster, using one of the three tests described before (line 9). If the test's result indicates a low variance between the sets then, the algorithm considers this cluster as a final cluster and it puts it in the list of final clusters (lines 9, 10 and 11). Else, it divides the initial cluster in K sub clusters by applying K-means algorithm (line 13).

To improve the performance of the K-means, we propose two values for K: $K_1 = 2$ or $K_2 = \lfloor\sqrt{n/2}\rfloor$, to divide dynamically each time a cluster to K sub clusters. The first value is a logic value that divides a cluster containing sets with high variance among them into two sub clusters, while the second value is defined in the *rule of thumb* [33]. This procedure of dividing clusters iterates until all sub clusters converge to the low variance between their sets (Fig. 3).

Algorithm 3 K-means Adapted to Variance Study.

Require: Set of measures' sets $M = \{M_1, M_2 ... M_n\}$, K.
Ensure: Set of clusters $C = \{C_1, C_2 ... C_K\}$.
1: $C \leftarrow \emptyset$ // list of all final clusters
2: $Q \leftarrow \emptyset$ // a temporary list of clusters
3: $C_1 \leftarrow \emptyset$
4: **for** each set $M_j \in M$ **do**
5: $C_1 \leftarrow C_1 \cup \{M_j\}$
6: **end for**
7: $Q \leftarrow Q \cup \{C_1\}$
8: **repeat**
9: **if** Variance Study(C_i) is true **then**
10: $C \leftarrow C \cup \{C_i\}$
11: remove C_i from Q
12: **else**
13: $Q \leftarrow Q \cup$ K-means(C_i, K)
14: **end if**
15: **until** no cluster $C_i \in Q$
16: return C



(a) $K = K_1 = 2$ (b) $K = K_2 = \lfloor\sqrt{n/2}\rfloor$

Fig. 3. Searching clusters using improved K-means.

Subsequently, we propose a different manner for each test to calculate the Euclidean distance between the centroids of clusters and a set (e.g. $d(M_j, r_i)$ in Algorithm 2). Since Fisher and Tukey tests are based on the measures inside the sets and the means of the sets when calculating the result R (Equations 1, 2 and 3), we propose to calculate the Euclidean distance between the mean of a set and the centroids of clusters when assigning this set to the nearest centroid in the K-means algorithm. In other word, K-means computes first the means of all sets then it selects randomly K means of sets to be the initial centroids of clusters and it begins to assign each set to the nearest centroid based on $d_m(\overline{Y_j}, r_i)$ in definition 5. On the other hand, the Euclidean distance between the variance of a set and the centroids, i.e. $d_v(\overline{Y_j}, r_i)$ in definition 6, is calculated when using the Bartlett test since its result is based on the variance of sets (Equations 4 and 5).

B. Redundancy Deleting at the CH

After having identified the final clusters that contain redundant data sets, the CH deletes redundancy from each cluster in order to reduce the amount of data transmitted to

the sink. Algorithm 4 shows how the CH selects the data sets to be sent to the sink among redundant sets in each cluster. Instead of sending all the data sets in each cluster, the CH decides to send only one useful information to the sink which corresponds to the data set with the highest number of measures.

Algorithm 4 Selecting Sets Algorithm.

Require: Set of clusters $C = \{C_1, C_2, \ldots, C_K\}$.
Ensure: List of selected sets, L.
1: $L \leftarrow \emptyset$
2: **for** each cluster $C_i \in C$ **do**
3: consider the set M_j has the longest cardinality in C_i, (i.e., $|M_j| > |M_{j*}|$; where $M_{j*} \in C_i$)
4: $L \leftarrow L \cup \{M_j\}$
5: **end for**
6: return L

VI. Experimental Results

In this section, we present the experimental results of our technique, at both sensor node and CH levels. We show via simulation on real data the efficiency of our approach in saving energy and reducing the huge amount of data thus extending the network lifetime of real UASNs. We used real data collected from the Argo project [34]. Argo deployed more than 3000 sensors distributed over the global oceans which collect salinity and temperature measurements from the upper $2000m$ of depth. In this paper, we are interested in 240 sensors deployed in the Indian ocean over an area of $5000 \times 5000m^2$. Then, these nodes are classified into three clusters of $N_1 = 40, N_2 = 80$ and $N_3 = 120$ sensors respectively. Each node reads periodically real measures while applying the first aggregation phase. At the end of this step, each node sends its set of measures with frequencies to their corresponding CH which in his turn applies the CH aggregation phase. For the sake of simplicity, in this paper we are interested in one field of sensor measurements: the salinity[2]. Furthermore, we compare our results to those obtained by applying the Prefix Frequency Filtering (PFF) technique used for periodic sensor networks [20]. We evaluated the performance using the following parameters: **a)** δ, which defines the threshold for $Similar$ function between two measurements. We varied δ to : $0.01, 0.03$ and 0.05. **b)** $\mathcal{T}$, the number of sensor measurements taken by each sensor node during a period. We varied $\mathcal{T}$ to: 200, 500 and 1000. **c)** α, the false-rejection probability in the ANOVA model which we varied to 0.01 and 0.05. and **d)** K, number of sub clusters resulting from K-means wich takes two values $K_1 = 2$ and $K_2 = \sqrt{n/2}$.

A. Data aggregation ratio at the sensor node

Due to the "$Similar$" function, each sensor node has the ability to reduce the amount of data collected at each period by eliminating redundancy from them. Fig. 4 shows the percentage of remaining data which will be sent to the CH without and with applying the local aggregation phase at the nodes level. The obtained results show that each sensor node will send in the worst case scenario, e.g. $\delta = 0.01$ and $\mathcal{T}$=200, 24% of its collected data to the CH after applying the aggregation phase comparing to 100% of its collected data without applying it. These results are very interesting in terms of eliminating redundancy from data sent by each sensor to the CH. Subsequently, we can observe that the aggregation phase eliminates more data redundancy when δ or $\mathcal{T}$ increases, because $Similar$ function will find more similar measures at each period.

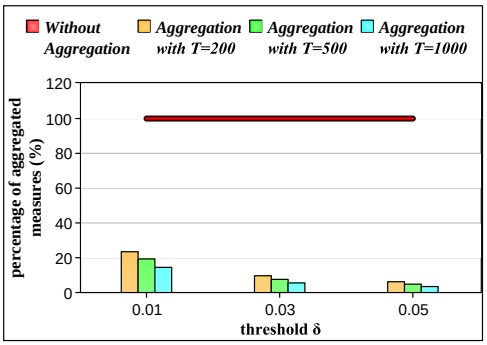


Fig. 4. Percentage of aggregated measures at the sensor nodes.

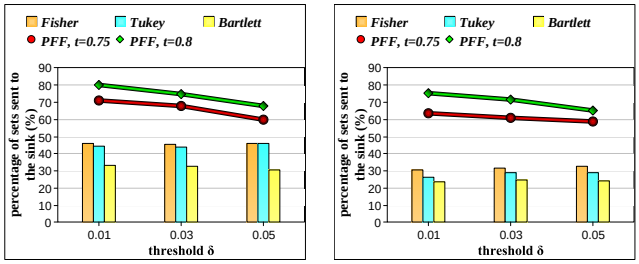
B. Data aggregation ratio at the CH

In this section, we show how each CH is able to reduce the redundancy among sets sent by its sensor nodes at each period before sending them to the sink (Fig. 5). First, we fixed in Fig. 5 (a, b and c) $\mathcal{T}$ and risk α and we varied the number of sensor nodes (N) for a CH to N_1, N_2 and N_3 respectively. Then, we fixed N and α in Fig. 5 (d, b and e) and we varied $\mathcal{T}$ to 200, 500 and 1000 respectively. After that, we fixed N and $\mathcal{T}$ in Fig. 5 (b and f) and we varied α to 0.01 and 0.05 respectively. The obtained results show clearly that, our technique allows CH to eliminate, when varying δ, more redundant sets at each period comparing to the PFF technique. This is because, the variance condition used in the ANOVA model is more flexible, in terms of finding redundant sets, comparing to the Jaccard similarity function used in PFF. We can also observe that, CH can reduce 25 to 66% of data sets sent to the sink comparing to the PFF, with different tests and parameters used.

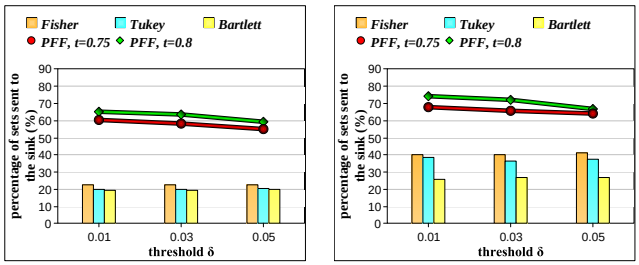
Several observations can be made based on the results in Fig. 5:

- Bartlett test sends the less percentage of sets to the sink comparing to Fisher and Tukey tests. This is because Bartlett test is more flexible regarding the variance between measures (Equations 4 and 5) compared to the variance calculated in Fisher (Equation 1) and Tukey (Equations 2 and 3).
- CH eliminates more redundant sets when N increases (subfigs. a, b and c). This is because, when a CH has more sensor members, their data, collected over the given area, will be more similar.

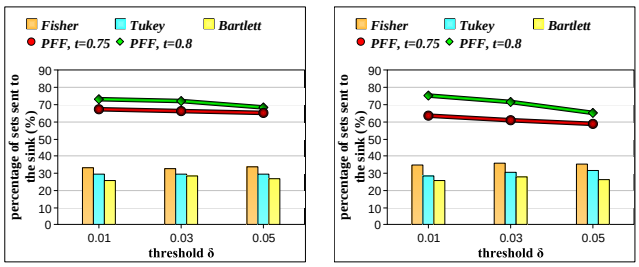
[2]the other is done by the same manner.



(a) $N = N_1, \mathcal{T} = 500, \alpha = 0.01$ (b) $N = N_2, \mathcal{T} = 500, \alpha = 0.01$



(c) $N = N_3, \mathcal{T} = 500, \alpha = 0.01$ (d) $N = N_2, \mathcal{T} = 200, \alpha = 0.01$



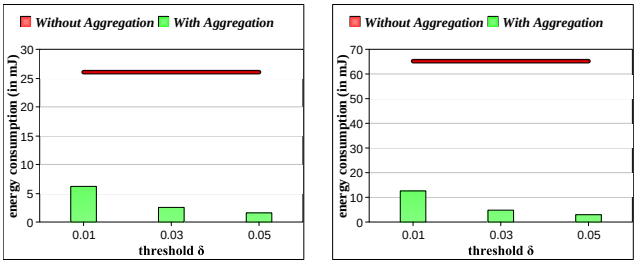
(e) $N = N_2, \mathcal{T} = 1000, \alpha = 0.01$ (f) $N = N_2, \mathcal{T} = 500, \alpha = 0.05$

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

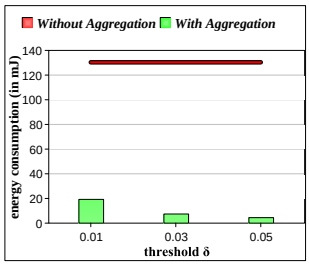
- The percentage of sets sent to the sink for the three tests is almost fixed when increasing δ in each subfigure. This is because, the data set saves the same variance when changing δ.
- CH eliminates more redundant sets in the three tests when decreasing α (subfigs. b and f). This is because, when the risk α increases the null hypothesis will a have higher probability of being rejected.

C. Energy consumption study

In this section, our objective is to study the energy consumption at the sensor nodes and CH levels. Therefore, energy consumption in sensor networks is highly dependent on amount of data sent and received. First, Fig. 6 shows the energy consumption comparison with and without applying the aggregation phase by each sensor node and when varying $\mathcal{T}$ and δ. Since the aggregation phase reduces significantly the redundancy among data collected by the sensor node (see Fig. 4), it allows it to save proportionally its energy when transmitting its data to the CH at each period. This truth is clearly shown in Fig. 6 when the sensor node applies the aggregation phase and when δ or $\mathcal{T}$ increases. It is important to notice that our technique can conserve energy of a sensor node up to 96%.



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



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

Fig. 6. Energy consumption at each sensor node.

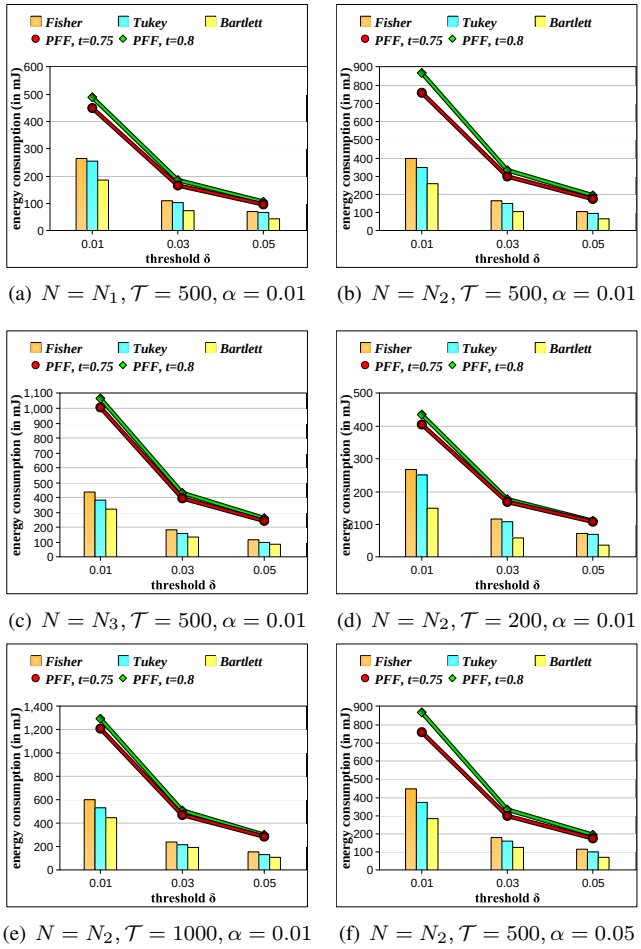
On the other hand, Fig. 7 shows the energy consumption comparison between the ANOVA model with the three tests and the PFF technique at the CHs level. In Fig. 7 (a, b and c), we fixed $\mathcal{T}$ and α and we varied N to N_1, N_2 and N_3 respectively, while we fixed N and α in Fig. 7 (d, b and e) and we varied $\mathcal{T}$ to 200, 500 and 1000 respectively. Then, we fixed N and $\mathcal{T}$ in Fig. 7 (b and f) and we varied α to 0.01 and 0.05 respectively. The obtained results show that our technique outperforms PFF for all values of thresholds and it reduces up to 70% of the energy consumption when compared to PFF. This result is logical since CHs eliminate more sets when using the ANOVA model comparing to the PFF technique (see Fig. 5).

Therefore, the above truth allows us to conclude some observations shown in Fig. 7:

- Bartlett test decreases energy consumption of the CHs more than the other tests.
- CHs conserve more energy comparing to PFF when N increases (subfigs. a, b and c).
- The energy consumption at the CHs is more minimized when α decreases (subfigs. b and f).

D. K-means study

In this section, we show first how many times each CH will apply the K-means algorithm on the received sets at each period, in order to obtain final clusters that contain redundant sets. Then we will show the total number of iterations generated when applying K-means at each period. The obtained results are dependent on the number of member nodes N of a CH. They are also dependent on how many sub clusters the CH divides each time an intermediate cluster that contains sets with high variance, e.g. K_1 or K_2, and from the risk α used in ANOVA model. We show, in this section, the obtained results when $\mathcal{T}$ is fixed to 500 measures. First, Fig. 8 (a, b

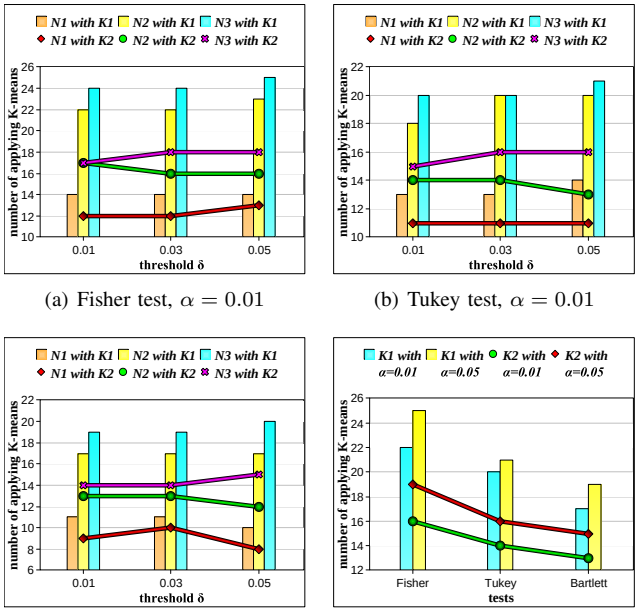


(a) $N = N_1, \mathcal{T} = 500, \alpha = 0.01$ (b) $N = N_2, \mathcal{T} = 500, \alpha = 0.01$

(c) $N = N_3, \mathcal{T} = 500, \alpha = 0.01$ (d) $N = N_2, \mathcal{T} = 200, \alpha = 0.01$

(e) $N = N_2, \mathcal{T} = 1000, \alpha = 0.01$ (f) $N = N_2, \mathcal{T} = 500, \alpha = 0.05$

Fig. 7. Energy consumption at the CHs.



(a) Fisher test, $\alpha = 0.01$ (b) Tukey test, $\alpha = 0.01$

(c) Bartlett test, $\alpha = 0.01$ (d) $N = N_2, \delta = 0.03$

Fig. 8. Number of applying K-means by the CHs at each period, $\mathcal{T} = 500$.

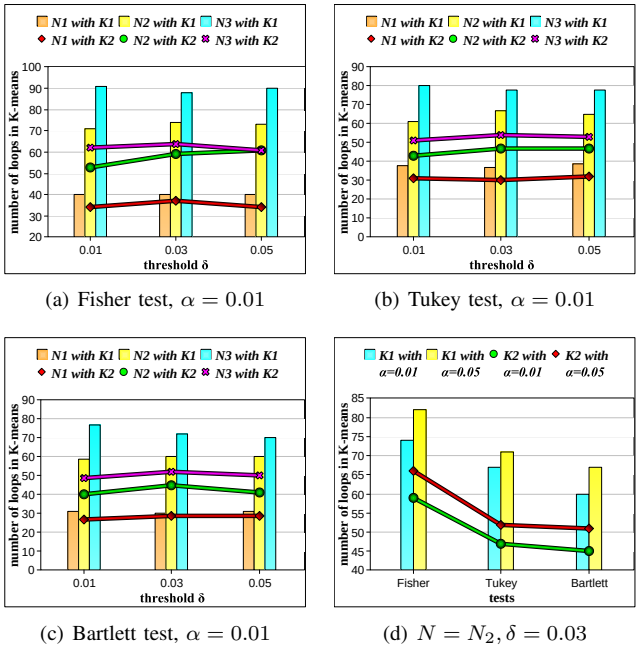
and c) shows the number of applying K-means by the CHs at each period with the three ANOVA tests when fixing α to 0.01 and varying K to K_1 and K_2. Contrarily, we fixed in Fig. 8 (d) N and δ and we varied α to 0.01 and 0.05. The obtained results show that the CH applies K-means few times. Applying K-means in such a way is very effective since it finds, dynamically, the optimal number of clusters without any initialization at the beginning. Also, several observations can be made based on the results of Fig. 8:

- Bartlett is the best test in terms of applying K-means. It quickly finds the final clusters compared to applying it with other tests because of its flexibility regarding the variance between measures (see observations for Fig. 5).
- The number of applying K-means increases when N increases (Fig. 8 (a or b or c)). Obviously, when a CH has more member nodes it must apply more K-means to find the final clusters.
- Dividing an intermediate cluster into K_2 sub clusters when applying K-means is a more flexible way to find final clusters comparing to apply it with K_1, for the same value of α.
- CHs apply less number of K-means when α decreases. This is because when α decreases the number of redundant sets at each final cluster increases (see Fig. 5).

Fig. 9 shows the total number of iterative loops when each CH applies K-means at each period. This factor is very important since it can prove, on the one hand, that the procedure always terminates and on other hand it can affect the latency of the aggregation phase at the CH. The number of iterative loops in K-means is significantly related to the initial selected cluster centroids which we tacked randomly in our case. We varied the parameters similarly to Fig. 8 with the three tests of ANOVA. The obtained results show the CH needs a small number of loops at each period to apply K-means, with the different used parameters. For instance, in the worst case scenario, when $N = N_3$ with $K = K_1$ and $\delta = 0.01$ with $\alpha = 0.01$ in Fisher test the CH needs 91 loops to apply 24 times K-means (see Fig. 8 (a)) consequently in each time it needs only 4 loops. Therefore, K-means algorithm seems very suitable for the computation resources in the CH. Therefore, similar observations to the results of Fig. 8 can be made regarding the number of iterative loops in K-means with the different values of parameters. This is because, the number of loops is highly dependent on the number of applied K-means and the random selection of the cluster centroids.

VII. Conclusion and Future Work

Although the research on acoustic underwater sensor networks has significantly advanced in the last decades, energy consumption still remains the major challenge to be optimized. We propose in this paper a two-tier data aggregation based transmission-efficient technique for periodic UASN which applies at each cluster separately in a clustering network. The node member aims to eliminate redundancy from data collected at each period at the first tier. Then, CH applies K-means algorithm adopted to the one-way ANOVA model with three statistics tests in order to eliminate redundancy from members that generate redundant data sets. The experiment



(a) Fisher test, $\alpha = 0.01$ (b) Tukey test, $\alpha = 0.01$

(c) Bartlett test, $\alpha = 0.01$ (d) $N = N_2, \delta = 0.03$

Fig. 9. Number of iterative loops in K-means at each period, $\mathcal{T} = 500$.

results show that our technique has largely reduced the data redundancy in the whole network and has also extended the network lifetime.

As a future work, we plan to schedule the sensor nodes in the network in a manner that nodes generating redundant data will not be active at the same time. Thus, sensor nodes will conserve more energy and network lifetime will be extended.

REFERENCES

[1] J. Heidemann, M. Stojanovic and M. Zorzi, *Underwater sensor networks: applications, advances and challenges*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 370, pp. 158-175, 2012.

[2] C. Peach and A. Yarali, *An Overview of Underwater Sensor Networks*, The Ninth Int. Conf. on Wireless and Mobile Communications (ICWMC 2013), pp. 31-36, July 2013.

[3] I. Akyildiz, D. Pompili and T. Melodia, *Underwater acoustic sensor networks: research challenges*, Journal of Ad Hoc Networks, Vol. 3, No. 3, pp. 257-279, May 2005.

[4] N. Javaid, M. R. Jafri, Z. A. Khan, N. Alrajeh, M. Imran and A. Vasilakos, *Chain-Based Communication in Cylindrical Underwater Wireless Sensor Networks*, Journal of Sensors 2015, Vol. 15, No. 2, pp. 3625-3649, February 2015.

[5] S. Sendra, J. Lloret, J. J.P.C. Rodrigues and J.M. Aguiar, *Underwater wireless communications in freshwater at 2.4 GHz*, Communications Letters, IEEE, Vol. 17, Iss. 9, pp. 1794-1797, 2013.

[6] M. Garcia, S. Sendra, M. Atenas and J. Lloret, *Underwater wireless ad-hoc networks: A survey*, Mobile ad hoc networks: Current status and future trends, pp. 379-411, 2011.

[7] K. T.-M. Tran and S.-H. Oh, *UWSNs: A Round-Based Clustering Scheme for Data Redundancy Resolve*, International Journal of Distributed Sensor Networks, Vol. 2014, 6 pages, 2014.

[8] M. Garcia, S. Sendra, J. Lloret and R. Lacuesta, *Saving energy with cooperative group-based wireless sensor networks*, In Cooperative Design, Visualization, and Engineering, Springer Berlin Heidelber, pp. 73-76, 2010.

[9] G. Yang, M. Xiao, E. Cheng and J. Zhang, *A cluster-head selection scheme for underwater acoustic sensor networks*, in Proc. of the 2010 International Conference on Communications and Mobile Computing (CMC'10), pp. 188-191, April 2010.

[10] A. Anbarasan, S. Sivasubramaniam, M. Mohanasundhram, *A Minimum Cost Effective Cluster Algorithm Using UWSN*, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, Iss. 7, July 2014.

[11] J. Lloret, M. Garcia, J. Tomas and J.J.P.C. Rodrigues, *Architecture and protocol for intercloud communication*, Information Sciences, Vol. 258, pp. 434-451, 2014.

[12] R. Kumar and N. Singh, *A Survey on Data Aggregation And Clustering Schemes in Underwater Sensor Networks*, International Journal of Grid Distribution Computing, Vol. 7, No. 6, pp. 29-52, 2014.

[13] K. T.-M. Tran and S.-H. Oh, *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.

[14] N. Goyal, M. Dave, A.K. Verma, *Fuzzy based clustering and aggregation technique for Under Water Wireless Sensor Networks*, International Conference on Electronics and Communication Systems (ICECS), pp. 1-5, 2014.

[15] X. Kui, J. Wang, S. Zhang and J. Cao, *Energy Balanced Clustering Data Collection Based on Dominating Set in Wireless Sensor Networks*, Ad Hoc & Sensor Wireless Networks Journal, Vol. 24, No. (3-4), pp. 199-217, 2015.

[16] J. Lloret, M. Garcia, D. Bri and J.R. Diaz, *A Cluster-Based Architecture to Structure the Topology of Parallel Wireless Sensor Networks*, Sensors, Vol. 9, No. 12, pp. 10513-10544, 2009.

[17] Z. Liu, J. Wang, S. Zhang, H. Liu and X. Zhang, *A Cluster-Based False Data Filtering Scheme in Wireless Sensor Networks*, Adhoc & Sensor Wireless Networks, Vol. 23, Iss. (1-2), pp. 41-45, 2014.

[18] B. Zhao and V. Friderikos, *Increased Energy Efficiency via Delay-Tolerant Transmissions in Cognitive Radio Networks*, Journal of Network Protocols and Algorithms, Vol. 5, No. 2, pp. 31-49, 2013.

[19] M.Y. Mowafi, F.H. Awad, M.A. Al-Batati, *Opportunistic Network Coding for Real-Time Transmission over Wireless Networks*, Network Protocols and Algorithms, Vol. 5, Iss. 1, pp. 1-19, 2013.

[20] J. Bahi, A. Makhoul, and M. Medlej, *A Two Tiers Data Aggregation Scheme for Periodic Sensor Networks*, Ad Hoc & Sensor Wireless Networks, Vol. 21, No. (1-2), pp. 77-100, 2014.

[21] H. Harb, A. Makhoul, R. Tawil and A. Jaber, *A Suffix-Based Enhanced Technique for Data Aggregation in Periodic Sensor Networks*, 10th IEEE Int. Wireless Communications and Mobile Computing Conference (IWCMC 2014), p. 494-499, 2014.

[22] K. T.-M. Tran, S.-H. Oh and J.-Y. Byun, *Well-suited similarity functions for data aggregation in cluster-based underwater wireless sensor networks*, International Journal of Distributed Sensor Networks 2013, Article ID 645243, 7 pages, 2013.

[23] D. Laiymani and A. Makhoul, *Adaptive data collection approach for periodic sensor networks*, In 9th International Wireless Communications and Mobile Computing Conference (IWCMC), pp.1448-1453,2013.

[24] R. Hall, *http://web.mst.edu/psyworld/tukeysexample.htm*, Psychology World, 1998.

[25] H. Arsham and M. Lovric, *Bartlett's Test*,The International Encyclopedia of Statistical Science, Springer, Part 2, pp. 87-88, 2011.

[26] H. Harb, A. Makhoul, D. Laiymani, A. Jaber and R. Tawil, *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.

[27] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, in Proceedings of the Fifth Berkeley Symposium, Vol. 1, pp. 281-297, 1967.

[28] S. Wang, F. Dai and B. Liang, *A path-based clustering algorithm of partition*, Information and Control, Vol. 40, Iss. 1, pp. 141-144, 2011.

[29] J. Wu, X. Li, T. Sun and W. Li, *A density-based clustering algorithm concerning neighborhood balance*, Journal of Computer Research and Development, Vol. 47, Iss. 6, pp. 1044-1052, 2010.

[30] C. Zhang and Z. Fang, *An Improved K-means Clustering Algorithm*, Journal of Information & Computational Science, Vol. 10, No. 1, pp. 193-199, 2013.

[31] B. M, *Choosing the number of clusters*, Data Mining and Knowledge Discovery, Vol. 1, Iss. 3, pp. 252-260, May/June 2011.

[32] A. Ahmed and W. Ashour, *An Initialization Method for the K-means Algorithm using RNN and Coupling Degree*, International Journal of Computer Applications, Vol. 25, No. 1, July 2011.

[33] K. Mardia, J. Kent, J. Bibby, *Multivariate Analysis*, Academic Press Editor, 1979.

[34] Argo project, *http://www.argo.ucsd.edu/index.html*.