Energy-efficient adaptive distributed data collection method for periodic sensor networks

Ali Kadhum M. Al-Qurabat

Department of Software,
University of Babylon,
Babylon, Iraq
Email: alik.m.alqurabat@uobabylon.edu.iq

Ali Kadhum Idrees*

Department of Computer Science,
University of Babylon,
Babylon, Iraq
Email: ali.idrees@uobabylon.edu.iq
*Corresponding authors

Abstract: This article suggests a method, called energy-efficient adaptive distributed data collection method (EADiDaC), which collects periodically sensor readings and prolong the lifetime of a periodic sensor network (PSN). The lifetime of EADiDaC method is divided into cycles. Each cycle is composed of four stages. First, data collection. Second, dimensionality reduction using adaptive piecewise constant approximation (APCA) technique. Third, frequency reduction using symbolic aggregate approximation (SAX) approach. Fourth, sampling rate adaptation based dynamic time warping (DTW) similarity. EADiDaC allows each sensor to remove the redundant collected data and adapts its sampling rate in accordance with the monitored environment conditions. The simulation experiments on real sensor data by applying OMNeT++ simulator explains the effectiveness of the EADiDaC method in comparison with two other existing methods.

Keywords: periodic sensor networks; PSNs; data collection; adaptive sampling rate; adaptive piecewise constant approximation; APCA; dynamic time warping; DTW similarity; symbolic aggregate approximation; SAX; network lifetime.

Reference to this paper should be made as follows: Al-Qurabat, A.K.M. and Idrees, A.K. (xxxx) 'Energy-efficient adaptive distributed data collection method for periodic sensor networks', *Int. J. Internet Technology and Secured Transactions*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Ali Kadhum M. Al-Qurabat is currently a PhD student in the Software Department, University of Babylon, Iraq. He received his MSc degree in IT from Universiti Tenaga Nasional (UNITEN), Malaysia, in 2012. He received his BSc degree in Computer Science from University of Babylon, Iraq, in 2002. He joined the Department of Computer Science, University of Babylon, Iraq, in 2006, where he is currently an Assistant Lecturer. His research interests include e-procurement, sensor networks, WSN, data aggregation.

Ali Kadhum Idrees received his BSc and MSc degrees in Computer Science from the University of Babylon, Iraq in 2000 and 2003 respectively. He received his PhD in Computer Science (Wireless Networks) in 2015 from the University of Franche-Comte (UFC), France. He is currently an Assistant Professor in Computer Science at the University of Babylon, Iraq. His research interests include wireless networks, wireless sensor networks, distributed computing, data mining, evolutionary and swarm intelligent algorithms, and optimization in communication networks.

1 Introduction

The future of internet will include a huge number of interconnected nodes expressing various things from tiny sensor nodes and portable devices to large web servers and supercomputer clusters. This type of a worldwide network is called the internet of things (IoT) (Jun et al., 2011; Li et al., 2013). The effective data collection techniques are necessary for IoT in order to gather and process the data at IoT nodes (Jun et al., 2011; Ulusoy et al., 2011). In IoT, every sensor node has low cost, power supply, a speed of processing, bandwidth, and memory capacity. Sensor nodes are spatially deployed in the region of interest in order to monitor the physical or environmental phenomena like temperature, humidity, light, pollution, pressure and sound. They collect the sensed data from the monitored environment, manipulate the data locally, and transmit them to the sink for further analysis. These sensor nodes work in a collaborative manner and constitute a wireless sensor network (WSN) (Akyildiz and Vuran, 2010; Idrees et al., 2014, 2015; Abdelaal et al., 2016). WSN represents one of the big contributors in the IoT because of their widespread use in many applications such as agricultural, healthcare, transportation, environment, industry, and military (Wang et al., 2012a; Idrees et al., 2016).

One of the most critical constraints of the sensor node is the battery life. Due to the environment or cost restrictions, it is difficult or impossible to change or recharge the sensor batteries (Abdelaal et al., 2016; Wang et al., 2012a). Thus, the sensor nodes are deployed with high density in order to enhance the network lifetime. In sensor node, the radio unit represents the principal source of energy consumption. Therefore, it is important to remove redundant sensed data before reporting them to the sink to save the energy and improve the lifetime of sensor node (Tang and Xu, 2008). It is necessary to take into consideration data capturing, communication, and routing problems in order to design energy-saving protocol for PSN. Data collection approaches determine the way of sensor's work in data collection and sending to the base station. Therefore, data collection represents the crucial function in PSNs (Campobello et al., 2016; Jon, 2016).

In WSN, the collection of data can be categorised into two models: time-driven and event-driven (Abdelaal et al., 2016; Jon, 2016). This work considers time-driven data collection model which known as periodic sensor networks (PSNs). In PSN, every sensor node transmits the sensed data of the monitored area to the sink periodically. Several PSNs applications use the periodic way to monitor certain conditions regularly such as pressure, humidity, temperature, etc. Two main challenges in PSN, first, PSN has to provide adequate lifetime in order to satisfy application's needs. Second, data management is more difficult due to the huge amount of collected data by this network.

The radio communication represents the most influential factor on energy consumption of PSNs. Therefore, sending a lot of data to the base station leads to various undesired issues such as degrading the quality of data, network congestion, and energy consumption (Wang et al., 2012a). Some proposed works focus on reducing communication overhead using data reduction techniques (Mohsenifard and Ghaffari, 2016). They reduce the amount of data communicated by every sensor nodes via use their processing capabilities to locally execute simple computations and transmit only required and partially processed data (Gupta, 2010). Data reduction aims to prolong the network lifetime and facilitate data analysis and decision making. In PSN, the change in the monitored environment can slow down or speed up. The energy consumption can be decreased when the sensor node modifies its sampling rate based on the dynamic modification of the monitored phenomena. Therefore, to prolong the network lifetime, adaptive sampling for periodic data collection is required for energy optimisation and data reduction (Gupta, 2010).

This paper introduces the following contributions.

1 A new method called EADiDaC is devised to collect the sensor data in an adaptive way such that the volume of data is reduced while PSN lifetime is enhanced. The principal idea of EADiDaC method is to utilise the similarity of collected data and adapts its sampling rate accordingly. EADiDaC works into cycles. Four stages in each cycle: gathering of data, dimensionality reduction using APCA technique, frequency reduction using symbolic aggregate approximation (SAX) technique, and adjusting the rate of sampling by using dynamic time warping (DTW) distance measure. The sensor node provides a new sampling rate after each cycle based on the similarity between the periods of one cycle.

2 A new adaptive sampling rate algorithm-based DTW similarity is suggested. In each cycle, the speed of readings capturing inside the sensor node depends mainly on the previously calculated sampling rate adaptively. EADiDaC method uses SAX approach to eliminate the redundancy in the collected measures before sending them the base station in order to conserve the energy and enhance the network lifetime.

3 The simulation results are accomplished by OMNeT++ network simulator to illustrate the effectiveness of the EADiDaC method. EADiDaC method has been compared to two algorithms in the related works: PFF algorithm that proposed by Bahi et al. (2014) and Harb et al. algorithm that introduced in Harb et al. (2016).

The rest of this paper is organised as follows. Next section exhibits literature review. Section 3 explains the description of EADiDaC method. Method evaluation is shown in Section 4. Finally, we present the conclusion and future works in Section 5.

2 Literature review

This section reviews some related literature concerning the adaptive data collection in WSNs. Adaptive collection approaches are considered as a good candidate to save energy and extend the network lifetime of PSNs. The major objective of an adaptive collection technique is to make the sensor node be able to change its sampling rate dynamically in accordance with the monitored environment conditions. This can reduce the repetitive gathered data, consume less energy, and decrease the processing load at the base station

(Gupta, 2010). Adaptive collection avoids capturing the redundant samples so as to reduce the volume of sent data to the base station and prolong the PSN lifetime.

In order to conserve the energy of a PSN, several MAC protocols are proposed (Nam et al., 2006; Van Dam and Langendoen, 2003; Zheng et al., 2005; Ye et al., 2002). The authors in Nam et al. (2006) proposed an adaptive MAC protocol to ensure the pre-configured network lifetime while the end-to-end latency is reduced. The protocol achieves this goal by using adaptive duty cycle which is adjusted using the pre-configured lifetime and the ratio of the remaining energy to the initial energy. Therefore, the sensor node with high energy wakes up repeatedly in order to deal with relaying data, whilst the sensor node with low energy becomes in sleep mode for a long time. The work in Zheng et al. (2005) presents a pattern-MAC (PMAC) protocol for WSN. It decides the schedules of sleep/wake up in an adaptive way for a node according to its own traffic, and the traffic patterns of its neighbours.

Adaptive collection avoids capturing the redundant samples by exploiting the correlation [temporal (Chatterjea and Havinga, 2008; Masoum et al., 2012), spatial (Willett et al., 2004; Wang et al., 2012b), or spatio-temporal (Masoum et al., 2013; Gedik et al., 2007; Liu et al., 2007)] between sensed data. The works proposed in Willett et al. (2004) and Wang et al. (2012b) considers adaptive sampling schemes-based spatial correlation among the physical sensed data. In Willett et al. (2004), the sampling rate is adapted by the base station. Initially, the base station activates a set of sensors to get the sensed data of monitored environment. The correlation percentage is computed for the received sensed data to increase or decrease the activated sensors. Some other approaches study temporal correlation among sensed data (Chatterjea and Havinga, 2008; Masoum et al., 2012). Chatterjea and Havinga (2008) present a sampling algorithm-based temporal correlation among sensed data. In this algorithm, the sampling rate is modified depending on the stability of the monitored environment. The sampling rate increases when the environment conditions are unstable, otherwise the rate decreases. Spatio-temporal correlation is used by some adaptive sampling techniques such as in Liu et al. (2007), Gedik et al. (2007) and Masoum et al. (2013). For instance, Masoum et al. (2013) introduce an energy-saving mechanism for data collection. Their scheme exploits spatio-temporal correlation among sensors and their sensed data to determine the candidate sensors which are responsible for sampling and transmission. The selected sensors are adaptively changed.

Some researchers used prediction techniques as a way to adjust the sampling rate of sensor nodes and to conserve energy of PSN (Jain and Chang, 2004; Alippi et al., 2010; Liu et al., 2005; Law et al., 2009; Padhy et al., 2010; Lazaridis and Mehrotra, 2003; Le Borgne et al., 2007; Tulone and Madden, 2006; Jain et al., 2004). An energy saving information gathering scheme is proposed by Liu et al. (2005) to predict the sampling rate inside sensor using ARIMA model. In Law et al. (2009), the authors presented an algorithm for adaptive sampling using Box-Jenkins approach to estimate the future sensor readings, depending on the existing readings. Alippi et al. (2010) introduced a power aware adaptive sampling method for snow monitoring. Their algorithm provides online estimation based on fast Fourier transform. An adaptive sampling method-based Kalman filter is introduced in Jain et al. (2004). This method used Kalman filter in both of base station and sensor nodes to predict future samples. When the sensor nodes filters fail to predict a future sample within a bounded precision, the base station will receive an update in order to update its filter accordingly.

The authors in Mehrnoush et al. (2015), Layuan et al. (2007), de Graaf (2013), Jun et al. (2008) and Huang et al. (2013) investigated the impact of the proposed protocols behaviours on the network performance. These protocols deal with some protocol evaluation metrics to improve the lifetime of the network, throughput, connectivity, etc.

In recent years, several adaptive sampling approaches in PSNs have been studied (Makhoul et al., 2015; Srbinovski et al., 2015; Laiymani and Makhoul, 2013; Zhang et al., 2015; Bahi et al., 2014; Harb et al., 2016). Laiymani and Makhoul (2013) proposed a scheme for adaptive sampling using ANOVA model and Fisher test in PSNs. This algorithm works at the sensor node to adapt its sampling rate. The authors in Bahi et al. (2014) proposed a method to remove the repetition of collected data in PSN called prefix frequency filtering (PFF). Makhoul et al. (2015) suggested adaptive data gathering approach for PSN. They combine between ANOVA model and remaining energy to permit every sensor node to modify its sampling rate in accordance with environment dynamics. Srbinovski et al. (2015) proposed a power saving data collection algorithm for power scavenging in WSNs. Their approach takes the energy harvesting from the monitored sensing area and modifies its sampling rate based on the remaining energy and observed environment. An adaptive sampling algorithm based on an endocrine regulation mechanism (EASA) in WSN is presented in Zhang et al. (2015). The EASA algorithm uses hormone information to control the nodes in working state or resting state and adjusts collecting frequency dynamically. Harb et al. (2016) proposed adaptive data collection approach based set similarity among sensor readings. Their technique allows each sensor node to identify, first, the similarity between data collected among successive periods using set similarity function, then to adjust its sampling rate to the newly calculated score of similarity. The sensor node reduces the amount of redundant collected readings and extends the network lifetime.

This paper proposed an energy-efficient adaptive distributed data collection (EADiDaC) method for PSNs. The main objective of EADiDaC is to remove redundant sensor readings, save energy, and improve the network lifetime. EADiDaC performs four main phases. First, data collection according to the adaptive sampling rate. Second, adaptive piecewise constant approximation (APCA) is applied to reduce the dimensionality of the collected sensed data. Third, SAX technique is used to remove the redundancy in the collected data and then transmits them to the sink. Fourth, EADiDaC allows to each sensor node to adapts its sampling rate for each cycle (cycle = 2 periods) based on the DTW similarity. EADiDaC is simulated on the OMNeT++ network simulator using real data of sensor nodes. The comparison results show that EADiDaC method can provide a better performance and prolong the network lifetime.

3 Description of the EADiDaC method

The description of EADiDaC method is given in more details in this section. The primary goal of EADiDaC method is to enable every sensor node to adjust its sampling rate adaptively according to the dynamic changing in the monitored environment. As a result, the redundancy in the collected readings will be reduced, and the consumption of energy will be minimised (prolong the lifetime of PSN), while the quality of collected readings is maintained sufficiently to permit significant analysis. Figure 1 illustrates the flowchart of the proposed EADiDaC method. This section describes in detail EADiDaC method stages and algorithms associated with each stage. Table 1 explains some parameters used in this paper.

Figure 1 Flowchart of proposed EADiDaC method (see online version for colours)

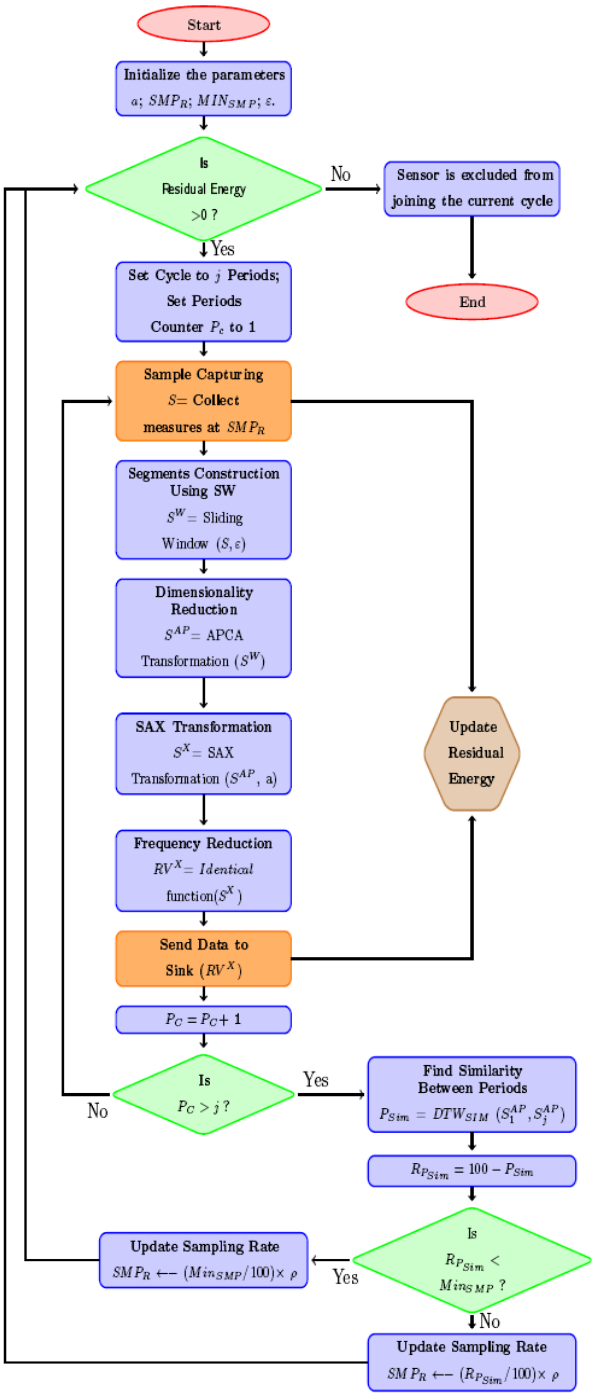


Table 1 Some parameters used in this paper

SMP_R	Sampling rate
MIN_{SMP}	Application criticality
S	Temperature readings series $S = s_1,\ldots,s_n$
S^{AP}	APCA of S, $S^{AP} = c_1^{ap},\ldots,c_w^{ap}$
S^{sx}	Symbolic representation of S^{AP}, $S^x = c_1^x,\ldots,c_w^x$
ε	Reconstruction error bound
a	Number of alphabet (for instance, if the alphabet = (w, x, y, z), $a = 4$)
β	Breakpoints, $\beta = \beta_1,\ldots,\beta_{a-1}$
n	Sensor id
n_e	Remaining energy of sensor n

3.1 Data collection

The sensor network consists of N sensor nodes (n_1, n_2,…,n_N) and a base station node. EADiDaC method is a periodic and works into cycles. The cycle includes two periods ($j = 2$). The period is partitioned into time slots. Therefore, each sensor node n captures the data reading periodically and at a specific speed (SMP_R). Consequently, the time-ordered sequence of sensed data constitutes a time series, $S_i = \{s_1, s_2,\ldots,s_{\rho-1}, s_\rho\}$, where ρ is the total number of temperature readings generated by sensor node n_i every T seconds. Therefore, EADiDaC method treats the sensor readings as a time series and named it as a temperature readings series. The SMP_R is initiated to ρ temperature readings per period. The redundant temperature readings captured by the sensor node increase in two states: short time slot and slowly variation of a monitored area of interest.

3.2 Dimensionality reduction

Since time series representation has a great impact on the simplicity and effectiveness of data readings mining; therefore, it is required to choose the suitable technique to represent the sensed readings series (Cassisi et al., 2012). Several representation methods are found in the literature such as discrete Fourier transform (DFT), the discrete wavelet transform (DWT), and singular value decomposition (SVD) (Cassisi et al., 2012). EADiDaC method uses a simple and efficient representation technique called APCA (Wang et al., 2013; Chakrabarti et al., 2002).

Normally, the aim of deploying sensor nodes is to measure the region of interest at fixed periods, and this produces a time-ordered series of samples which constitute a temperature readings series. Often, the volume of temperature readings series is very huge. Therefore, it is not practical to send all the collected readings from every sensor node back to the base station due to the constrained bandwidth and energy consumption on data sending. The dimensionality ρ of temperature readings series (which is the number of observed measures) have a direct proportionality relation with the communication cost. Thus, a smaller ρ can result in a significant reduction on the communication cost and hence, it will prolong the lifetime of the sensor network (Liu et al., 2007). In this stage, the EADiDaC method aims to minimise the volume of data of

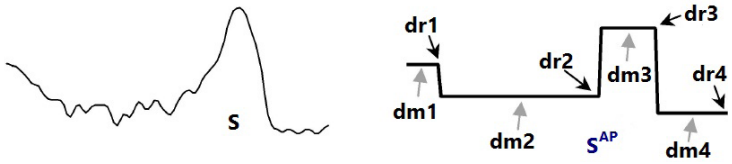
each sensor node before sending them to base station. Therefore, EADiDaC method achieves this task by using dimensionality reduction technique APCA.

In this stage, EADiDaC method transforms the temperature readings series $S = \{s_1, s_2,\ldots,s_{\rho-1}, s_\rho\}$ that collected during the first stage with SMP_R speed to an APCA representation in order to decrease the dimensionality of series. Data sorting is an integral part of data analysis. It improves the search and merges the sequences efficiently. Therefore, the efficiency of APCA is improved by sorting the sensed temperature readings in descending order so as to group the similar (or close similar) readings together. The APCA representation of S is given as follow

$$S^{AP} = \left\{\left\langle dm_1,\ dr_1\right\rangle,\ldots,\ \left\langle dm_m, dr_m\right\rangle\right\},\ dr_0 = 0 \tag{1}$$

The APCA divides the sorted temperature readings series S into a set of constant value segments (with a bounded reconstruction error ε) of varying lengths based on data such that their individual reconstruction errors are minimal. More formally, $|R(S^{AP}) - S| < \varepsilon$, $R(S^{AP})$ is the reconstruction function, and ε is an error threshold. Long segments are used to represent data regions of low activity, and short segments are used to represent regions of high activity (Zifan et al., 2007). Figure 2 illustrates this notation.

Figure 2 A temperature readings series S and its APCA representation S^{AP}, with $m = 4$ (see online version for colours)



The APCA approximates each segment S_j^{AP} by a pair (dm_j, dr_j) of two numbers, where dm_j is the mean value of temperature readings in the j^{th} segment which is defined as

$$dm_j = \frac{\sum_{k=dr_{j-1}+1}^{dr_j} s_k}{dr_j - dr_{j-1}}. \tag{2}$$

Whilst dr_j is the right endpoint of the j^{th} segment (Wang et al., 2013).

By using the standard form of APCA with constant number of segments of varying lengths can influence on the accuracy of temperature readings. Hence, the problem addressed here is: for a given temperature readings series S and a given reconstruction error bound ε, find the number of segments to approximate the time series, such that the difference between any approximation value and its actual value is less than ε.

In our method, we make some slight modifications on APCA. First, the number of segments m will not be constant and determined priori, but it will be adaptive based on the user specified reconstruction error ε. In order to achieve this goal (i.e., making the number of segments adaptive), the sliding window algorithm is utilised. The reason for making the number of segments adaptive is to increase the accuracy of approximated measures by using a user specified reconstruction error. Second, we modified dr to represent the length of the segments rather than record the locations of their right endpoints.

3.2.1 Sliding window algorithm

Several applications such as weather, medical, and stocks employ the algorithm of sliding window. It is a temporal approximation over the actual value of the time series data (Yahmed et al., 2015). At the end of each period, EADiDaC method will apply the sliding window algorithm on the collected readings to produce a different number of segments with varying lengths. The mechanism of sliding window algorithm is given as follows:

1 for the potential segment, tying its left point with the first point of a temperature readings series

2 try to approximate the data to the right side with expanding the long of segments

3 in a specific point x of the temperature readings series, the potential segment will has a reconstruction error exceed the user-specified threshold ε

4 hence, the subsequence from the tying point to $x - 1$ is converted into a segment.

5 tying point is put on the location x, and repeat steps 1, 2, 3, and 4 until all the temperature readings series has been converted into segments.

The sliding window algorithm is attractive because of its great simplicity, intuitiveness and particularly it is an online algorithm (Yahmed et al., 2015). Algorithm 1 represents the process of segment construction using sliding window algorithm.

Algorithm 1 Segments construction using sliding window

Require: S: ρ-dimensional temperature readings series, ε: reconstruction error bound.

Ensure: S^W: the set of segments with m subsets.

```
1  S ← Sorting(S)                          // Sorting temperature readings in descending order
2  Flag ← 1                                // Starting point
3  SEG_No. ← 1                             // Number of segments
4  while (x< ρ) do
5      x ← 2
6      while (Calculate_Error(S[Flag: Flag + x]) < ε) do
7          x ← x + 1
8      end while
9      S^W[SEG_No.] ← Create_Segment(S[Flag: Flag + x – 1])
10     Flag← Flag + x
11     SEG_No. ← SEG_No. + 1
12 end while
13 return S^W
```

Let S_i^W be the subset consisting of all the temperature readings on this segment $\{s_i, s_{i+1},\ldots,s_j\}$, which meet the reconstruction error bound such that the difference between the approximation value and the actual value is not larger than a given reconstruction error bound. Eventually, we have a set S^W of m subsets, where $S^W = (S_1^W, S_2^W,\ldots, S_m^W)$. After segmenting the temperature readings series using sliding window algorithm, the produced

set of segments S^W is used by Algorithm 2 to produce the APCA representation for temperature readings series S. Algorithm 2 illustrates the process of dimensionality reduction using APCA.

Algorithm 2 Dimensionality reduction using APCA

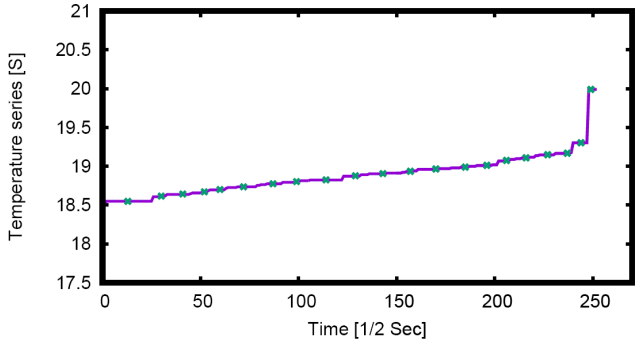
Require: S^W: the set of segments with m subsets

Ensure: S^{AP}: the set of segments with m subsets and two numbers per segment

```
for i ← 1 to m do
    SG ← S_i^W
    Sum ← 0
    Count ← 0
    for j ← 1 to Len(SG) do
        Sum ← Sum + SG[j]
        Count ← Count + 1
    end for
    SEG_len ← Count
    SEG_μ ← Sum / Count
    S_i^AP ← Create_segment(SEG_μ, SEG_len)
end for
return S^AP
```

Let $S_i^{AP}(SEG_{\mu_i}, SEG_{len_i})$ denote a subset consisting of all the temperature readings on this segment $\{s_i, s_{i+1},\ldots,s_j\}$, where SEG_{μ_i} is the mean of these temperature readings and SEG_{len_i} is the length of the segment. The problem mentioned above is solved by constructing a set of segments S^{AP} with m subsets $\{S_1^{AP}, S_2^{AP},\ldots,S_m^{AP}\}$, that meet the reconstruction error bound ε. Figure 3 is an example for the process of transforming 250 temperature readings into an APCA representation with 21 segments, where the reconstruction error ε is 0.1.

Figure 3 Example of APCA transformation (see online version for colours)



3.3 Frequency reduction

In PSN, radio communication is the most influential factor on energy consumption. Sending a lot of data to the base station leads to various undesired issues such as degrading the quality of data, network congestion, and energy consumption (Wang et al., 2012a). Furthermore, there is a directly proportional relationship between the communication cost and the dimensionality of temperature readings (Liu et al., 2007). Hence, in order to reduce the energy consumption and prolong the PSN lifetime while keeping up the acceptable accuracy of sent readings, the sensor node has to transmit to the base station as few as possible of sensed readings. The primary goal of this stage is to reduce the volume of temperature readings, which are assembled by every sensor node and keep the recurrence number of every reading in order to not impact on the readings analysis in the base station. In this stage, the EADiDaC method uses SAX symbolic algorithm (Lin and Li, 2009; Malinowski et al., 2013) as a method to remove the redundancy in the temperature readings series before sending them to the base station. SAX symbolic representation method is considered a pioneer in reducing dimensionality/numerosity of temperature readings series. It consists of two steps: piecewise aggregate approximation (PAA) transformation and the transformation of the numerical data into a set of symbols. Each symbol takes its value from a finite alphabet (Lin and Li, 2009; Malinowski et al., 2013). EADiDaC method uses SAX method because it requires a low processing cost and achieves high data reduction while keeping the primary features of temperature readings.

SAX transforms a temperature readings series S of length n into a reduced vector, for example, RV with length m. By applying the process to the series S = (8, 16, 6, 4, 2, 2, 2, 2, 2, 2, 20, 10) with length n = 12. The resulted reduced vector RV is (12, 5, 2, 2, 2, 15) of length m = 6. After that, the obtained reduced vector transformed into a symbolic based on the break points are determined by the Gaussian distribution as illustrated in Table 2. In this case, the produced vector is changed into the symbolic representation *CBAAAC*. In order to increase the efficiency of SAX method, the PAA method is replaced by APCA approach. The PAA method uses constant length segments whilst APCA approach allows creating segments with varying lengths. In the previously mentioned example, the values (2, 2, 2, 2, 2, 2) are decreased to (2, 2, 2), where it is aggregated to one value because of the constant length of m (Ganz et al., 2014). In EADiDaC method, we use a variable length m in order to get a smaller reduced vector when the readings have low activity. This achieves through replacing the PAA transformation by the APCA transformation. The APCA representation S^{AP} of the original temperature readings series S is transformed into SAX representation RV X using the following steps:

1 normalising temperature readings series to have zero mean and one standard deviation

2 partitioning the temperature readings series into an unspecified number of segments using sliding window algorithm with varying lengths

3 the mean (SEG_{μ}) and length (SEG_{len}) for each segment is computed

4 the mean values are quantised into symbols selected from an alphabet of size N.

Table 2 A lookup table of the breakpoints for *a*

	a	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
β_i	β_1	–0.43	–0.67	–0.84	–0.97	–1.07	–1.15	–1.22	–1.28
	β_2	0.43	0	–0.25	–0.43	–0.57	–0.67	–0.76	–0.84
	β_3		0.67	0.25	0	–0.18	–0.32	–0.43	–0.52
	β_4			0.84	0.43	0.18	0	–0.14	–0.25
	β_5				0.97	0.57	0.32	0.14	0
	β_6					1.07	0.67	0.43	0.25
	β_7						1.15	0.76	0.52
	β_8							1.22	0.84
	β_9								1.28

Steps 2 and 3 are APCA representation. In step 4, the quantisation uses (N – 1) breakpoints which partition the region under the Gaussian distribution into a equal proportional regions. Breakpoints can be defined as a sorted values list $B = \beta_1,\ldots,\beta_{a-1}$. The region under a $N(0, 1)$ Gaussian curve from β_i to $\beta_{i+1} = 1/a$, where β_0 and β_a refer to $-\infty$ and ∞ respectively. The breakpoints are located by search them in a statistical table. For instance, Table 2 shows a lookup table of the breakpoints for a with values range from 3 to 10 (Lin and Li, 2009; Malinowski et al., 2013).

When the breakpoints are determined, we can quantise the APCA coefficients as follow. Every APCA normalised mean value less than the smallest breakpoint will be converted to 'a' symbol, whilst the APCA normalised mean values that are equal to or larger than the smallest breakpoint and less than the second smallest breakpoint are converted into 'b' symbol, etc. For Figure 3, we use a = 5, SAX representation provides five symbols: a, b, c, d, and e. The symbols can be merged to introduce a sequence called word. It can be defined as follow. Let $alpha_i$ indicates the i^{th} value of the alphabet (i.e., $alpha_1$ = a and $alpha_2$ = b). Consequently, the transformation from a APCA representation S^{AP} to aword S^x is computed as follows

$$S_i^x = alpha_j, \quad iif \quad \beta_{j-1} \le s^{AP}\left(SEG_\mu\right)_i < \beta_j. \tag{3}$$

After converting the APCA mean values into SAX symbols, the resulted SAX symbols sequence will include redundant symbols due to multiple consecutive segments are transformed to the same symbol. In this stage, EADiDaC method removes these redundant symbols in each period to prevent transmitting the same symbols to the base station. Therefore, we will define a function that allows each sensor node to find the similarity among the symbols of $word\ S^x = s_1^x,\ldots,s_w^x$ to eliminate this redundancy. The identical function identifies the similarity between two symbols s_i^x and s_j^x and can be defined as follow

$$Identical\left(s_i^x, s_j^x\right) = \begin{cases} 1 & if \quad s_i^x = s_j^x \\ 0 & othewise \end{cases} \tag{4}$$

The sensor node n will search for the same symbols in the *word* of the period j. If the same symbols are found, the sensor will sum the means associated with each symbol. It

also sums the frequency associated with each symbol every time occurs in the word. After that, the average of accumulated means is found and the segment that contains (mean, frequency) is created. Otherwise, the sensor will create a new segment and add the mean and the associated frequency to it. The frequency of the symbol $Fr(s_i^x)$ is defined as the sum of the associated frequency of the same symbol in the same set.

3.4 Adaptive sampling rate

In this stage, EADiDaC method modifies its sampling rate based on the percentage of similarity between temperature readings of different periods in the cycle. The main purpose of this stage is to calculate the similarity among periods after each finished cycle to acclimate the rate of sampling according to the new similarity rate. EADiDaC method adapts its rate of sampling at the end of each cycle. Therefore, it uses the DTW distance measure to find the amount of similarity between periods of each cycle.

3.4.1 Similarity measure

The main purpose of using similarity measure is to exploit the similarity among periods in order to adjust the sampling rate according to the amount of similarity among periods for each cycle. At the end of each period, EADiDaC method uses the modified APCA technique on the collected measures to produce a different number of segments with varying lengths for each period. The Euclidean distance can not be used to calculate the distance between sequences whose lengths are different. Therefore, DTW distance measure has been adopted to overcome this problem (Cassisi et al., 2012). It is a widely used measure in data mining community. It is a utility for various tasks in time series problems including classification, clustering, and anomaly detection that allows time-axis scaling. The distance between two temperature readings series of varying lengths can be measured using DTW. It does not use one-to-one comparison such as in Euclidean but uses many-to-one (and vice versa) comparison (Cassisi et al., 2012). If we have two temperature readings series $Q = (q_1, q_2,\ldots,q_p)$ and $T = (t_1, t_2,\ldots,t_m)$ of length p and m respectively, a p-by-m distance matrix can be built in order to align the two sequences using DTW (Cassisi et al., 2012).

Algorithm 3 SAX frequency reduction

Require: S^{AP}: m subset of APCA coefficients, a: alphabet length, α: alphabetic

Ensure: RV^X: reduced vector of segments with two number per segment $\{SG_1(V_1, Fr_1),\ldots, SG_j(V_j, Fr_j)\}$, where $(j < m)$

1 **for** $i \leftarrow 1$ to m **do** //Normalise means of APCA coefficients

2 $Temp \leftarrow S_i^{AP}$

3 $D_{NOR} \leftarrow \frac{Temp[1] - \mu}{\sigma}$

4 $S_i^{AP} \leftarrow D_{NOR}$ $//S_i^{AP}(SEG_{\mu_i}, SEG_{len_i}, D_{NOR_i})$

5 **end for**

6 **for** $i \leftarrow 1$ to m **do** // quantised normalised mean of APCA into symbols

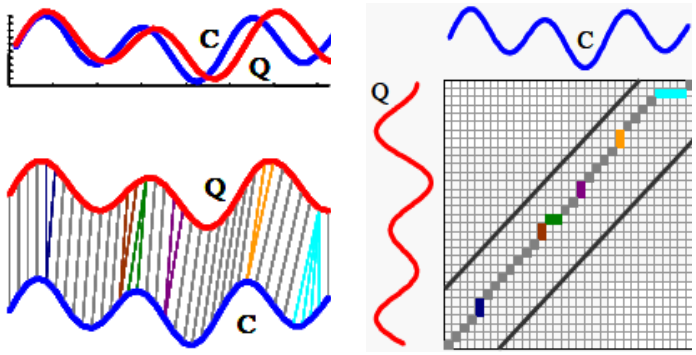
7 $Temp \leftarrow S_i^{AP}$

```
8       for j ← 1 to a do
9           if β_j ≤ Temp[3] < β_{j+1} then
10              S_i^AP ← α_j    //S_i^AP(SEG_μi, SEG_leni, D_NORi, α_j)
11          end if
12      end for
13  end for
14  z ← 0
15  for j ← 1 to a do
16      z ← z +1
17      Idx ← 0
18      for i ← 1 to m do
19          Temp ← S^APi
20          if Identical(Temp[4], α_j) = 1 then
21              Idx ← Idx +1
22              Fr_z ← Frz + Temp[2]
23              Total_μ ← Total_μ + Temp[1]
24          end if
25      end for
26      if Idx > 1 then
27          Vz ← Total_μ/Idx
28      end if
29      RV_z^X ← Create_segment(V_z, Fr_z)
30  end for
31  return RV^zX
```

$$DistMtrx(Q,T)=\begin{bmatrix} d(q_1,t_1) & d(q_1,t_2) & \cdots & d(q_1,t_m) \\ d(q_2,t_1) & d(q_2,t_2) & \cdots & d(q_2,t_m) \\ \vdots & \vdots & & \vdots \\ d(q_p,t_1) & d(q_p,t_2) & \cdots & d(q_p,t_m) \end{bmatrix}$$

where the element in the position (i^{th}, j^{th}) of the matrix contains the distance $d(q_i, t_j)$ between q_i and t_j. Usually the distance used in this matrix between two points is Euclidean distance $d(q_i, t_j) = (q_i - t_j)^2$. Each matrix element (i, j) corresponds to the alignment between the points q_i and t_j. This is illustrated in Figure 4.

Figure 4 (a) Two temperature reading series sequences (b) To align the sequences, we construct a warping matrix, and search for the optimal warping path (solid squares) (see online version for colours)



(a) (b)

The goal of DTW is to find the *warping path* $W = \{w_1, w_2,\ldots,w_k,\ldots,w_K\}$ of adjacent elements on *DistMtrx*, where $\max(p, m) \le K < p + m - 1$, and $w_k = DistMtrx(i, j)$, such that it minimises the following function

$$DTW(Q,T) = \min\left(\sqrt{\sum_{k=1}^{K}(w_k)}\right). \tag{5}$$

The warping path is ordinarily subject to few restrictions (Cassisi et al., 2012). Given $w_k = (i, j)$ and $w_{k-1} = (i', j')$ with $i, i' \le p$ and $j, j' \le m$:

1 *boundary conditions:* $w_1 = (1, 1)$ and $w_K = (p, m)$

2 *continuity:* $i - i' \le 1$ and $j - j' \le 1$

3 *monotonicity:* $i - i' \ge 0$ and $j - j' \ge 0$.

This path can be found using dynamic programming to evaluate the following recurrence which defines the cumulative distance matrix $\gamma(i, j)$ of the same dimension as the *DistMtrx*, where the distance $d(i, j)$ is found in the current cell and the minimum of the cumulative distances of the adjacent elements is

$$\gamma(i,\ j) = d\left(q_i, c_j\right) + \min\left\{\gamma(i-1, j-1), \gamma(i-1,\ j), \gamma(i,\ j-1)\right\}. \tag{6}$$

The last element of the warping path, w_K corresponds to the distance calculated with the DTW method (Cassisi et al., 2012). After finishing the distance calculation between two temperature readings series of the APCA representation Q_p and T_m, EADiDaC method uses *Similar* function to identify the similarity between them. The *Similar* function refers to the similarity between two APCA temperature readings series using the following formula

$$SIM\left(Q_p, T_m\right) = \frac{1}{1+\gamma(p, m)}. \tag{7}$$

After that, in order to measure the similarity percentage (P_{Sim}), the following formula can be used

$$P_{Sim} = \left(SIM\left(Q_p, T_m\right) \times 100 \right). \tag{8}$$

Algorithm 4 gives the similarity percentage (P_{Sim}) calculation between two APCA temperature readings series Q_p and T_m.

Algorithm 4 Similarity algorithm

Require: Two APCA temperature series Q_p and T_m

Ensure: P_{Sim}

```
for i ⇐ 1to len(Q) do
    for j ⇐ 1 to len(T) do
        Distance[i, j] ← (Q[i] – T[j])^2
    end for
end for
Accumulated_Cost[1, 1] ← Distance[1, 1]
for i ⇐ 1 to len(Q) do
    Accumulated_Cost[i, 1] ← Distance[i, 1] + Accumulated_Cost[i – 1, 1]
end for
for j ⇐ 1 to len(T ) do
    Accumulated_Cost[1, j] ← Distance[1, j] + Accumulated_Cost[1, j – 1]
end for
for i ⇐ 1 to len(Q) do
    for j ⇐ 1 to len(T) do
        Accumulated_Cost[i, j] ← Distance[i, j] + min(Accumulated_Cost[i, j – 1],
        Accumulated_Cost[i – 1, j], Accumulated_Cost[i – 1, j – 1])
    end for
end for
Sim ← 1 / (1 + Accumulated _ Cost[len(Q), len(T)])
P_Sim ← Sim × 100
return P_Sim
```

3.4.2 Verification the similarity of periods

In EADiDaC method, the sampling period refers to the time duration during which the sensor capture sensed temperature readings from the surrounding environment. The speed of change of environmental conditions and what fundamental features should be periodically gathered in temperature readings collection model can influence on the sampling period. In EADiDaC method, every node able to adapt its rate of sampling according to the amount of similarity among temperature readings series collected during different periods. The aim of computing the similarity between the temperature readings series every cycle is to adapt the rate of sampling based on the new calculated similarity.

Therefore, the APCA similarity coefficient is employed to discover the similarity percentage, P_{Sim} among several periods per cycle. On one hand, If P_{Sim} is high, it means the monitored condition is changed at a slow speed. Therefore, the sensor node will decrease its rate of sampling to the minimum value to prevent collecting redundant readings. On the other hand, if P_{Sim} is low, the sensor node will collect temperature readings at approximately maximum sampling rate so as to avoid missing significant measures. Therefore, to adapt the rate of sampling of sensor node in accordance with the computed similarity among periods, the reverse of similarity percentage for APCA similarity coefficient $R_{P_{Sim}}$ is computed as follow

$$R_{P_{Sim}} = 100 - P_{Sim}. \tag{9}$$

Consequently, the computed $R_{P_{Sim}}$ will be used to adapt the rate of sampling of the sensor in the new periods. When there is a high degree of similarity among periods (i.e., P_{Sim} is high), the sensor node balances its rate of sampling to the minimum value ($R_{P_{Sim}}$ is low). Otherwise, it balances its rate of sampling to the maximum value. As aforementioned, how the process of adapting the sampling rate in the sensor node depends on the $R_{P_{Sim}}$, the application criticality will be taken into consideration in this process.

3.4.3 Application criticality

The PSN can be used for monitoring disasters by using various kinds of sensor devices, e.g., for temperature, displacement, pressure, and concentration of chemicals, or noise detection. The influence of disasters on people and on the environment is not the same. Therefore, the sensor can modify its rate of sampling in a different manner for each monitored disasters. Therefore, if the risk level of the disaster is high then the sensor node must collect sensed readings more than if the risk level of the disaster is low. This can provide collected readings with high quality to make both the analysis easier and the monitored disaster is better to understand. There is an inversely proportional relation between P_{Sim} and $R_{P_{Sim}}$, therefore, when the similarity among periods is high, the $R_{P_{Sim}}$ will push the sensor node to make its sampling rate as minimum as possible.

In general, when the sensor node has the ability to alter its rate of sampling depending on the application's needs in PSNs, this will save its energy. In EADiDaC method, the criticality of application is expressed as a minimum amount of sampling rate in a period for a sensor node, MIN_{SMP}. MIN_{SMP} takes values in the range 0 to 100 which represent the criticality level either low or high respectively. The sensor node adapts the new sampling rate to the MIN_{SMP} and not to the $R_{P_{Sim}}$ when the recently calculated sampling rate is less than MIN_{SMP}. Depending on the requirements of the application and before the deployment, all the sensor nodes initialise their MIN_{SMP}. It is also possible to change MIN_{SMP} dynamically during the lifetime of the network for the whole sensors or for just a given subgroup of sensors if there are some types of management and control schemes are available.

Algorithm 5 illustrates an adaptive sampling rate approach. The main purpose of this algorithm is to give every sensor device the ability to modify its rate of sampling to conserve its power and to decrease the volume of collected data. Algorithm 5 works into cycles and each cycle consists of j periods. In each period, the sensor captures ρ temperature readings. The number of periods j is fixed to 2. For each cycle, the sensor

nodes look for the similarity percentage among periods (line 10), then it computes $R_{P_{Sim}}$ (line 11). Therefore, the sensor node will decide to increase its sampling rate to computed $R_{P_{Sim}}$ when it is greater than MIN_{SMP} which is determined by the application's needs. Otherwise, it decreases its sampling rate to the MIN_{SMP} (lines 12–16).

Algorithm 5 Adaptive sampling rate algorithm

Require: j (one cycle = j periods), ρ, MIN_{SMP}, a: alphabet, ε: reconstruction error

Ensure: SMP_R // new sampling rate

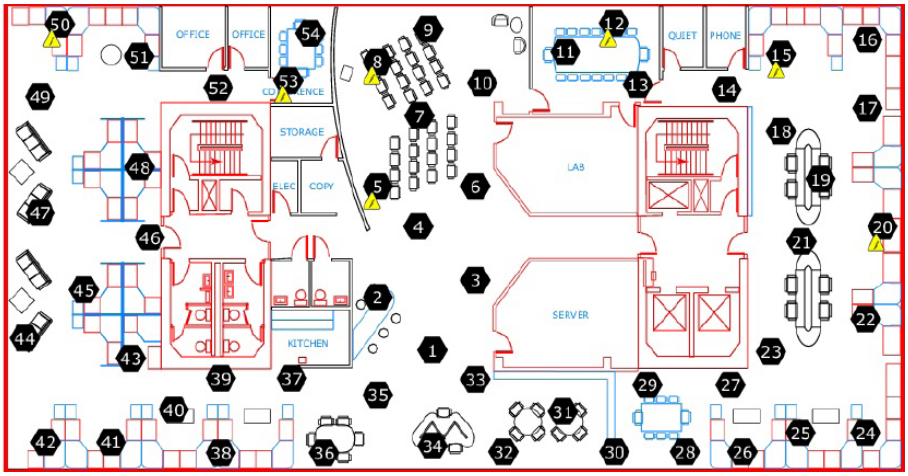
1 $SMP_R \leftarrow \rho$ //initialise sampling rate to ρ measures per period
2 **while** $n_e > 0$ **do**
3 **for** $i \leftarrow 1$ to j **do**
4 Collect temperature readings series (S_i)at SMP_R speed
5 $S_i^W \leftarrow$ segments construction using sliding Window (Si, ε)
6 $S_i^{AP} \leftarrow$ APCA dimensionality reduction (S_i^W)
7 $SendToSink(RV^X) \leftarrow$ SAX Frequency Reduction (S_i^{AP}, a)
8 **end for**
9 **for** each cycle **do**
10 $P_{Sim} \leftarrow$ Similarity (S_i^{AP}, S_j^{AP}), // S_i^{AP} the APCA coefficient set formed at period i
11 $R_{P_{Sim}} \leftarrow 100 - P_{Sim}$
12 **if** $R_{P_{Sim}} << MIN_{SMP}$ **then**
13 $SMP_R \leftarrow (MIN_{SMP}/100) \times \rho$
14 **else**
15 $SMP_R \leftarrow (R_{P_{Sim}}/100) \times \rho$
16 **end if**
17 **end for**
18 **end while**
19 return SMP_R

4 Method evaluation

4.1 Simulation framework

To study and evaluate EADiDaC method, extensive simulations are performed with discrete event simulator OMNeT++ (Varga, 2003). In these simulations, we consider N sensors deployed in the lab as illustrated in Figure 5. Sensors periodically capture local readings (e.g., temperature) at a specified rate. The base station is located at the centre of the lab. It receives sensed readings from each sensor node by a single hop.

Figure 5 Intel Berkeley lab sensor network (see online version for colours)



Source: Madden (2004)

EADiDaC method is distributed at each sensor node and it is based on the dataset of Intel Berkeley Research Lab (Madden, 2004). PSN in this Lab includes 54 Mica2Dot sensors localised as shown in Figure 5. The sensed data of the weather (such as temperature, humidity, and light) are periodically collected by these sensors once each 31 seconds. In our simulation, the sensor nodes use a log file contains about 2.3 million readings collected previously by Mica2Dot sensor nodes in the Lab. This article uses only one measure of sensor node measurements: temperature[1]. In Figure 5, every sensor node has a yellow sign is not used in our simulation because its data may be missed or truncated. Therefore, the temperature readings of 47 sensor nodes are selected and stored. The results are the average of 47 sensor nodes. Table 3 gives the selected parameters settings.

Table 3 simulation parameters for PSN initialisation

Parameter	*Value*
PSN size	47 nodes
a	5 and 10 symbols
ρ	20, 50, 100 and 200 readings
MIN_{SMP}	20, 40 and 60
ε	0.07, 0.1, 0.2 reconstruction error bound
j	2
E_{elec}	50 nJ/bit
β_{amp}	100 pJ/bit/m^2

In the experimental simulations, some performance metrics are applied to assess the effectiveness of the EADiDaC method such as sampling rate adaptation, number of collected temperature readings by a sensor node, number of sent temperature readings, energy consumption, accuracy, and lifetime. EADiDaC method uses the same energy

consumption model discussed in Harb et al. (2016). Energy consumed by the sensor node is just the periodically collected and sent temperature readings to the base station. The cost of transmission is calculated for a m – *bits* message and for a distance d as follow

$$E_{TX}(m,d) = E_{elec} * m + \beta_{amp} * m * d^2. \tag{10}$$

The energy consumption required for capturing m – *bits* by the sensor node is calculated as follow

$$E_{CX}(m,d) = E_{TX}(m,d)/7. \tag{11}$$

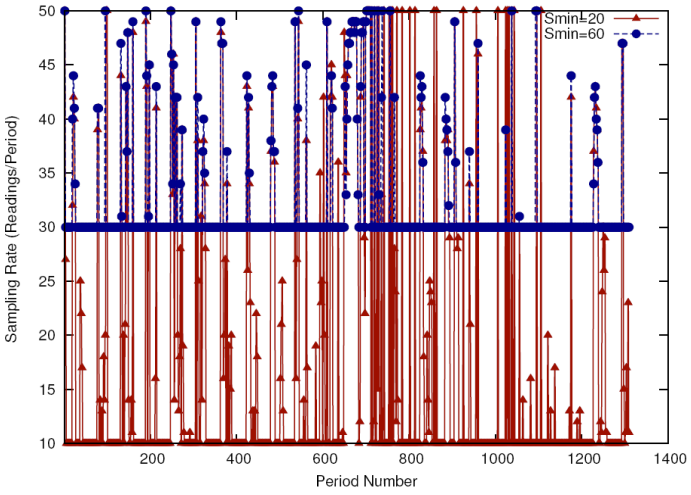
These experiment simulations consider the length of data reading m equal to 64. In the case of transmission, 16 bits are added to m – *bits* message which corresponds to the frequency of data reading m. Consequently, energy consumption is defined as the total energy dissipated at each sensor node during the collection and transmission of data readings and formulated as follow

$$E_{Total} = E_{TX}(m,d) + E_{CX}(m,d). \tag{12}$$

4.2 Performance analysis

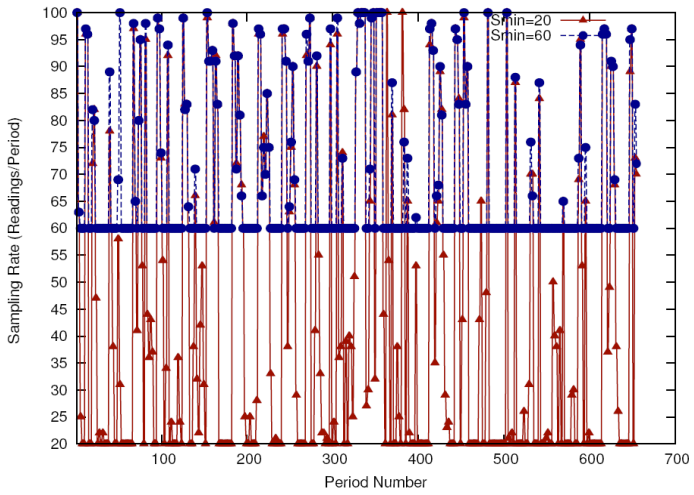
In this section, several experiments are achieved to show the performance of EADiDaC method. It is distributed at each sensor node in the PSN. Every node reads real temperature readings periodically and adapts its rate of sampling after each cycle based on the similarity percentage among collected sets of temperature readings.

Figure 6 Sampling rate adaptation, (a) ρ = 50 (b) ρ = 100 (see online version for colours)



(a)

Figure 6 Sampling rate adaptation, (a) $\rho = 50$ (b) $\rho = 100$ (continued) (see online version for colours)



(b)

4.2.1 Sampling rate adaptation

Figure 6 shows the adaptation of sampling rate and for two sizes of temperature readings (50 and 100 respectively), and reconstruction error bound ε is fixed to 0.1. The results illustrate the ability of sensor device to modify its rate of sampling dynamically depends on the application criticality level. The risk level MIN_{SMP} can be determined according to the type and requirement of application used to monitor the disaster.

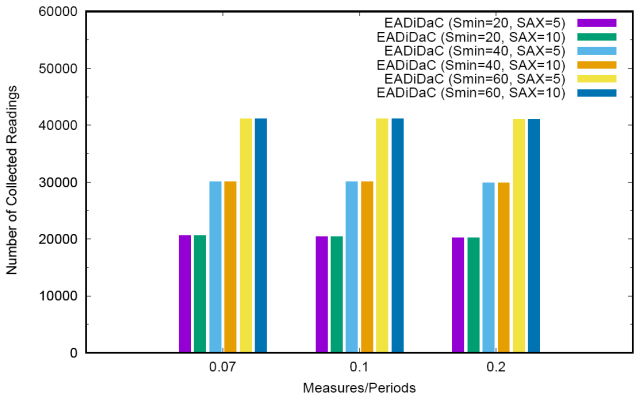
In this experiment, MIN_{SMP} uses two values: 20 for low risk level disaster and 60 for high risk level disaster. As shown in Figure 6, the adaptation of sampling rate is dynamic and after each cycle based on the application criticality level (i.e., $MIN_{SMP} = 20$ or 60). The results in Figures 6(a) and 6(b) validate the good performance of our method.

4.2.2 Number of collected readings

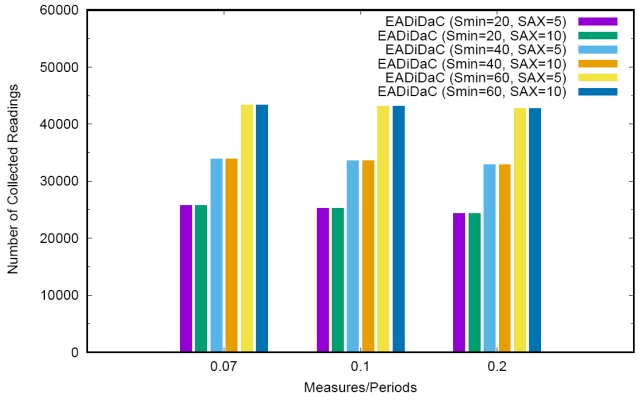
Figure 7 shows the number of collected readings by the node at the end of simulation. EADiDaC method uses different values for the parameters SMP_R, a, MIN_{SMP}, and ε.

As shown in these results, the alphabet size a does not affect the number of collected readings because of adaptation of sampling rate depends basically on the similarity among periods. EADiDaC method collects as large as possible of temperature readings when the MIN_{SMP} increases. This can support application requirements. When the risk level is high then EADiDaC method collects more readings. It can be seen that the increase in the ρ leads to increase the number of collected readings because of the decreasing the similarity percentage between collected readings of successive periods. It can be seen that when the reconstruction error bound ε increases then the number of collected readings decreases due to increasing the similarity among collected readings.

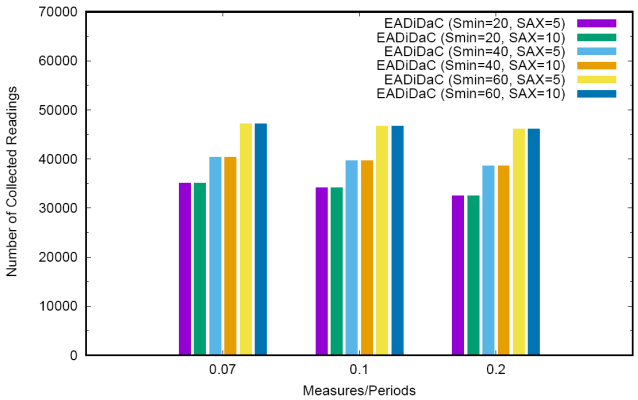
Figure 7 Number of collected readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)



(a)

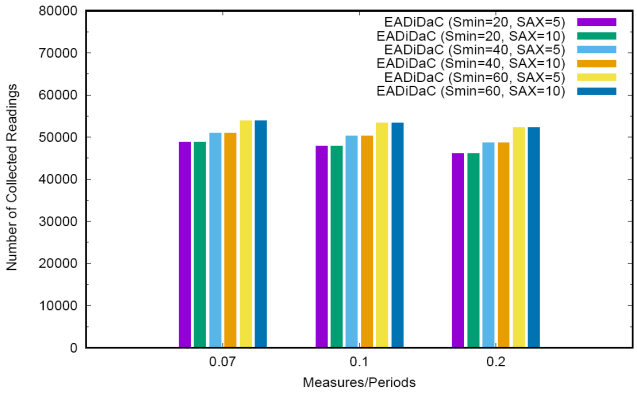


(b)



(c)

Figure 7 Number of collected readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



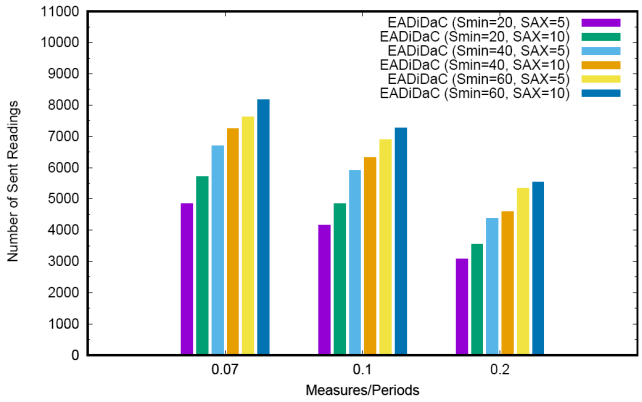
(d)

4.2.3 Number of sent readings

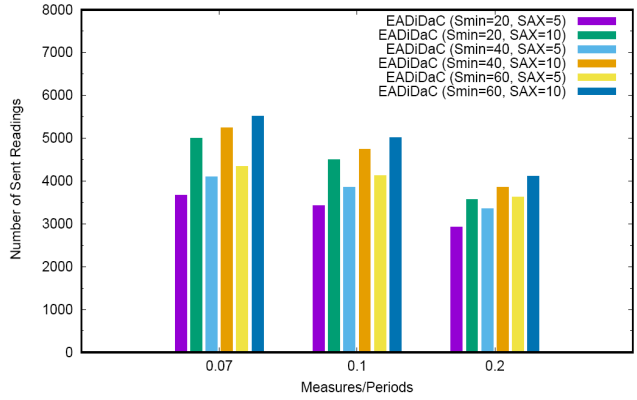
In this experiment, the number of sent readings by sensor node is evaluated. Another task carried out by EADiDaC method is to remove redundant collected readings before send them to the base station while maintaining the accuracy of collected readings. Figure 8 indicates the number of sent readings by the node at the end of the simulation.

Obviously, the number of sent readings increases with the number of alphabet sizes. This is due to the lack of similarity among collected readings. It can be seen that EADiDaC method send the larger amount of readings to the base station when the MIN_{SMP} increases or reconstruction error bound ε decreases. This can support the application's needs by sending a larger number of readings when the risk level of application is high. It is obvious that the increase in the SMP_R leads to decrease the number of sent measures due to the SAX technique that transforms the collected readings into fixed number of symbols, each one associated with different frequency. For example, suppose $a = 5$ and $SMP_R = 50$, the collected readings will be represented by five symbols (a, b, c, d, and e). Each of these symbols has a different associated frequency (e.g., 5, 4, 15, 10, 16). If the SMP_R increases to 100 for the same five symbols, it leads to represent the 100 collected readings by the same five symbols and with different associated frequency for each symbol. Therefore, EADiDaC method reduces the number of redundant data before send them to the base station to saves more energy and improve lifetime.

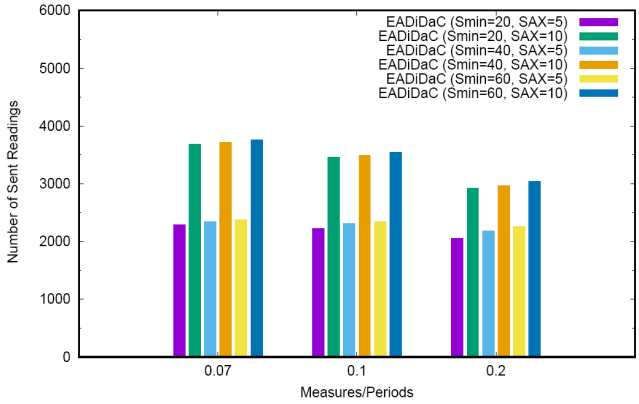
Figure 8 Number of sent readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)



(a)

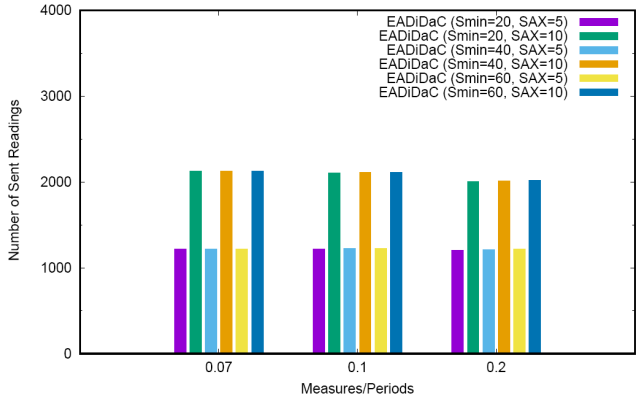


(b)



(c)

Figure 8 Number of sent readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)

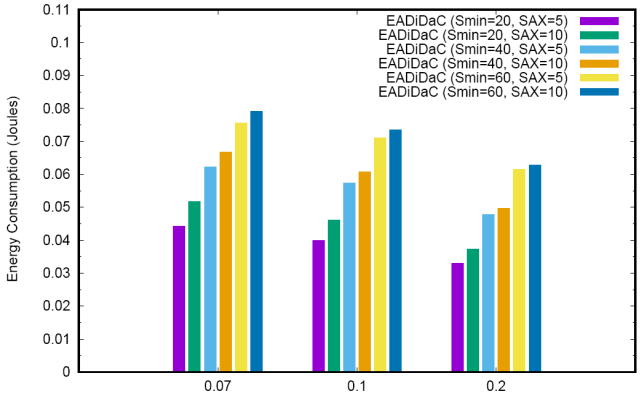


(d)

4.2.4 Energy consumption

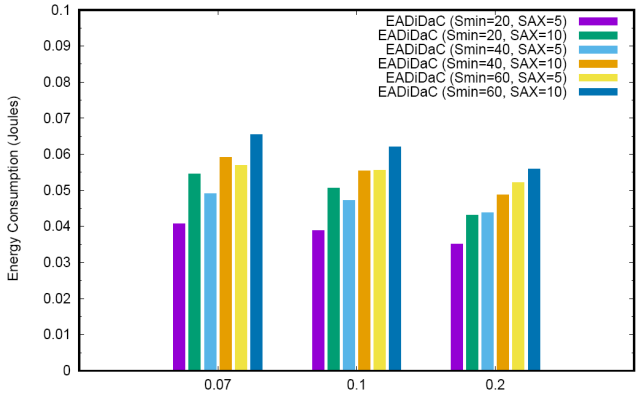
In this experiment, the energy consumption of the sensor node using EADiDaC method is studied. Figure 9 illustrates energy consumption by a sensor node at the end of the simulation. As shown in Figure 9, when the a increase, the number of sent readings increases (see Figure 8) thus energy consumption by the sensor node using EADiDaC method increases. EADiDaC method increases the sent readings when the risk level of the application is high. Therefore, the energy consumption by EADiDaC method increases when the MIN_{SMP} increases. Furthermore, it is obvious that the increase in the SMP_R or reconstruction error bound ε leads to decrease the number of sent readings thus save the energy of sensor node.

Figure 9 Energy consumption by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)

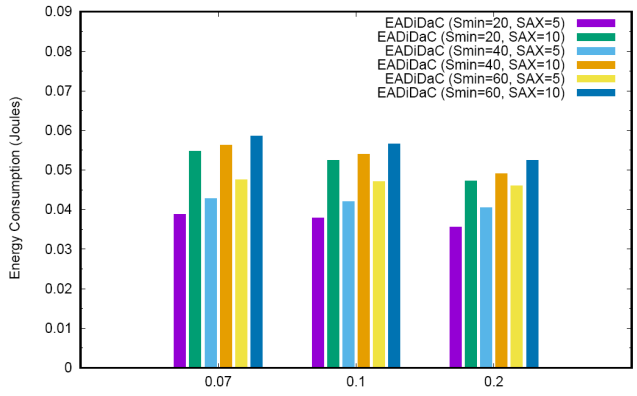


(a)

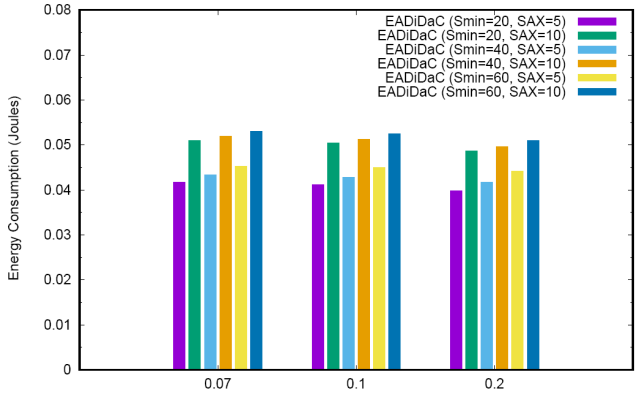
Figure 9 Energy consumption by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



(b)



(c)

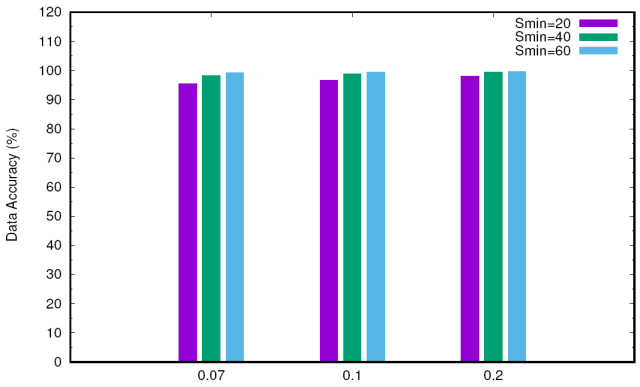


(d)

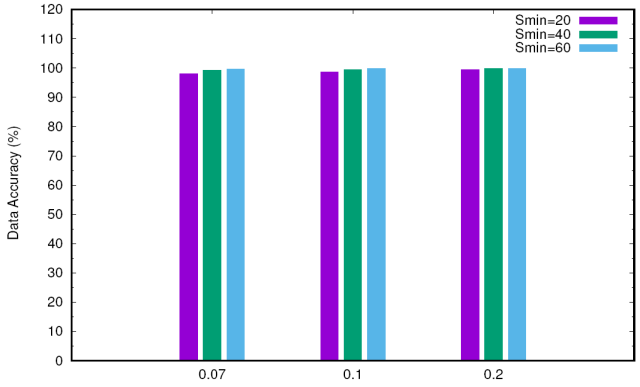
4.2.5 Data accuracy

Another important metric to evaluate the quality of EADiDaC method is the accuracy of collected data. It represents a measure of data loss rate. In order to evaluate the data accuracy, the lost data readings are counted in a periodic way after sampling rate adaptation of every sensor using our method. The data reading is considered as lost data reading if it collected by the sensor without adaptive sampling during the period p and it is not collected by the same sensor using our adaptive sampling technique for the same period. However, the data accuracy is computed at the end of the experiment by subtracting the lost data readings rate from the total number of data readings collected by the sensor without adaptive sampling. Figure 10 shows the data accuracy of EADiDaC method. The obtained results show that the EADiDaC method provides good performance regarding the data accuracy. It produces at least 95% of data readings accuracy. Therefore, the decision at the base station will be not influenced. Accordingly, EADiDaC method can be considered as an energy saving way to adapt the sampling rate of the sensor node while maintaining a high level of accuracy of the collected data readings.

Figure 10 Data accuracy of collected data readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)

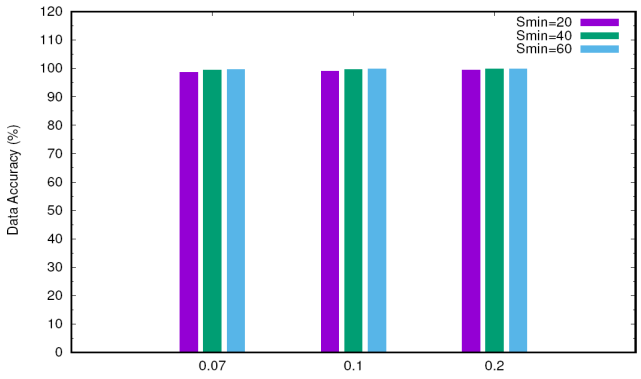


(a)

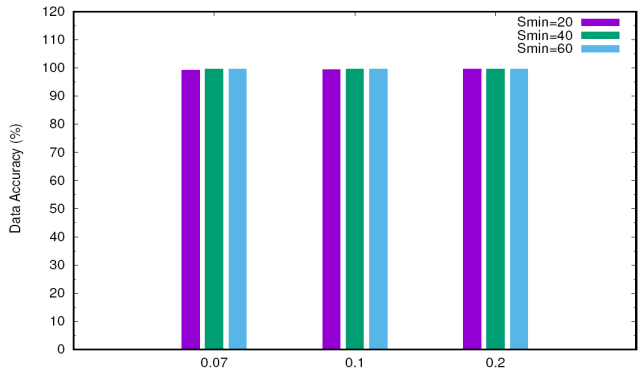


(b)

Figure 10 Data accuracy of collected data readings, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



(c)



(d)

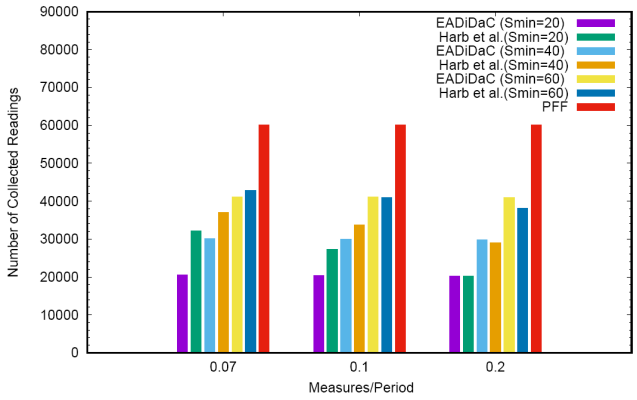
4.3 Comparison results

Depending on the conducted results in the Sub-section 4.2, EADiDaC method, with $a = 5$ seem to give the best results to be compared with the best results of other two existing techniques. The first scheme is called PFF that proposed by Bahi et al. (2014). The second approach is called Harb et al. that introduced in Harb et al. (2016).

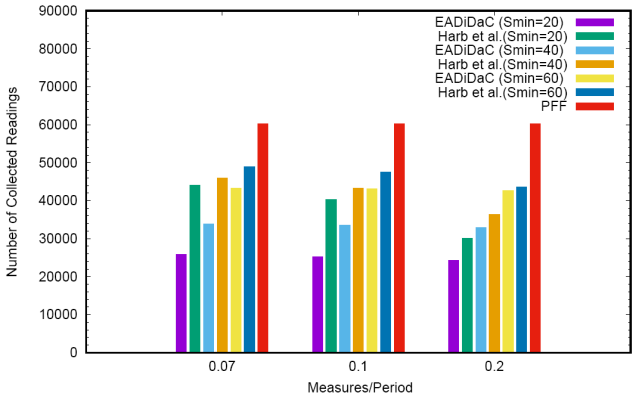
4.3.1 Number of collected readings

Figure 11 illustrates the number of collected readings at the end of simulation by every sensor node using EADiDaC method compared with other two approaches. EADiDaC method decreases the number of collected readings by a sensor node from 13% to 65% compared to PFF. The PFF does not allow to the sensor node to adapt its sampling rate therefore, it always collects the same number of readings. EADiDaC method decreases the collected readings from 1% to 35% in comparison with Harb et al. approach which allows the sensor node to adapt its rate of sampling based on the similarity between the periods of one cycle.

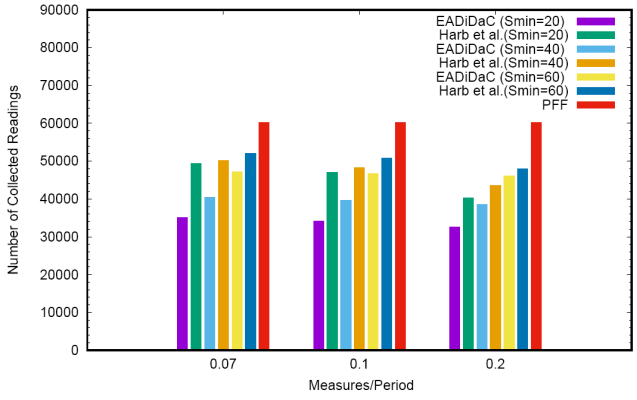
Figure 11 Number of collected readings by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)



(a)

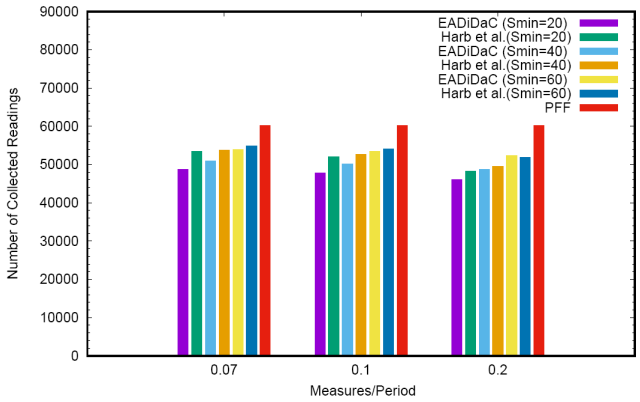


(b)



(c)

Figure 11 Number of collected readings by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



(d)

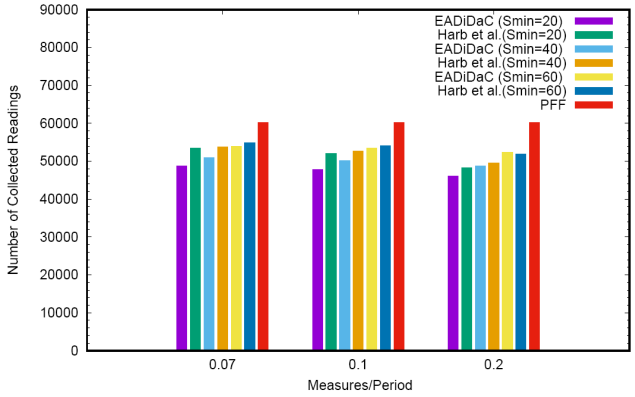
The results illustrate that EADiDaC method has the ability to get rid of the redundant collected readings efficiently so as to decrease the overhead of transmitted readings to the base station thus improve the network lifetime. It can be seen that EADiDaC method increases the volume of collected readings when the MIN_{SMP} is increased. This increment in the collected readings is to meet the application's requirements when the risk level is high.

4.3.2 Number of sent readings

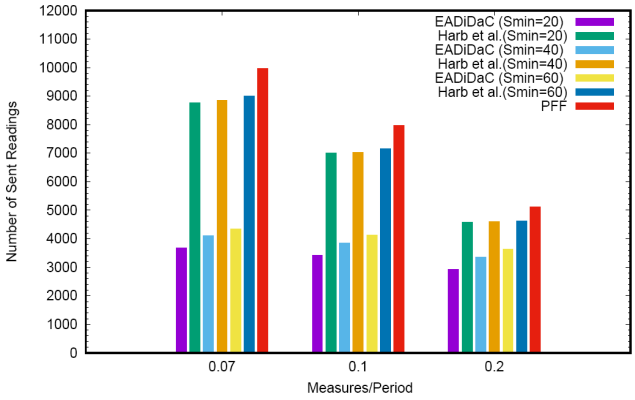
After collecting the data readings at each period, EADiDaC method at the sensor node able to decrease the number of sent readings to the base station by using SAX method. Therefore, EADiDaC method finds the redundant symbols in the *word* of each period and allocates for every symbol its frequency. Figure 12 demonstrates the number of sent readings by a sensor node to the base station at the end of simulation for EADiDaC method compared with the PFF and Harb et al. methods.

The results illustrate that EADiDaC method at each sensor node decreases up to 62% and 65% of the number of sent readings to the base station comparing to the PFF and 57% and 61% comparing to the Harb et al. methods respectively. Therefore, EADiDaC method removes the redundant collected readings successfully and the number of sent readings to the base station is reduced. We can also see that the volume of sent readings from the sensor node to the base station decreases when ρ increases or reconstruction error bound ε increases. This is due to the number of sent readings rely on the number of collected readings, ε, the identical function, and the risk level of application.

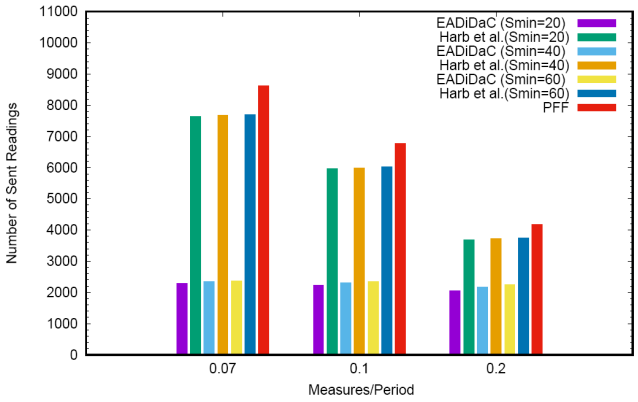
Figure 12 Number of sent readings by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)



(a)

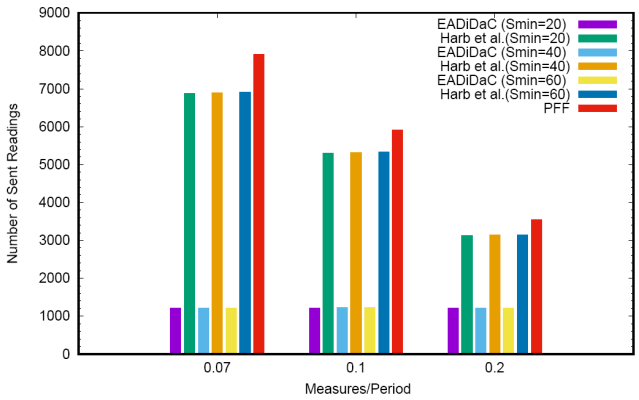


(b)



(c)

Figure 12 Number of sent readings by a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



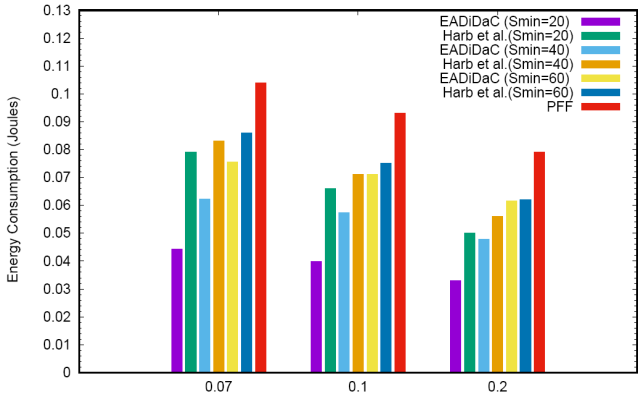
(d)

4.3.3 Energy consumption

Figure 13 shows the energy consuming by EADiDaC method at the sensor node compared with PFF and Harb et al. approaches.

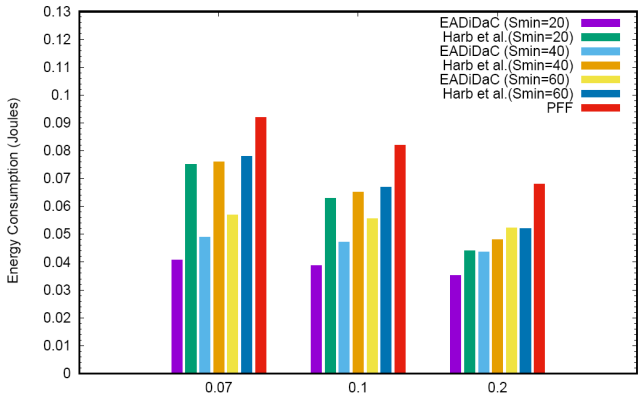
As shown in Figure 13, EADiDaC method outperforms the other approaches in term of energy consumption. It saves energy because it reduces both collected and sent readings at the sensor node. The consumed energy of a sensor node using EADiDaC method is minimised up to 57% and 27% compared to PFF and 43% and 15% compared to Harb et al. techniques respectively. It can be observed that EADiDaC method is effective in terms of reducing energy consumption for the applications with high and low risk level. It saves more energy when MIN_{SMP} is decreased.

Figure 13 Energy consumption, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)

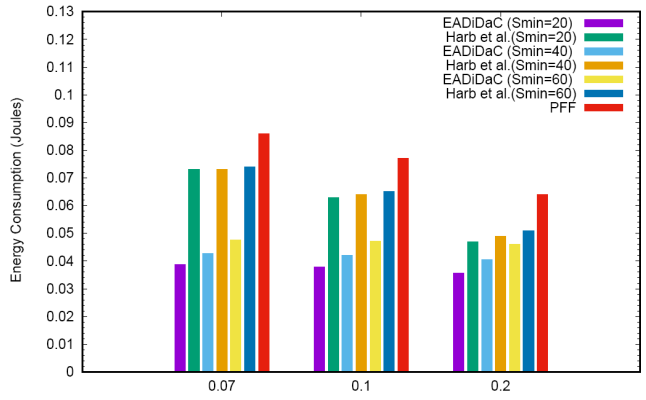


(a)

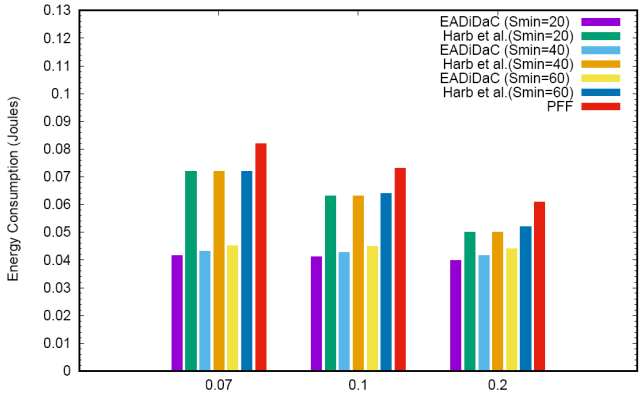
Figure 13 Energy consumption, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



(b)



(c)



(d)

4.3.4 Lifetime of sensor node

Finally, we study the influence of the number of collected and sent readings on the PSN lifetime. As exhibited by Figure 14, EADiDaC method gives a longer network lifetime compared with other approaches. Every sensor node initiated its energy to 40 mJ for the whole approaches in this comparison.

EADiDaC method enhances the lifetime of sensor node up to 55% compared to Harb et al. technique. These results are obtained due to the efficiency of EADiDaC method in conserving the energy of the sensor thus increases the PSN lifetime for both high and low risk level applications, whilst maintaining the quality of the gathered readings.

Figure 14 Lifetime of a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (see online version for colours)

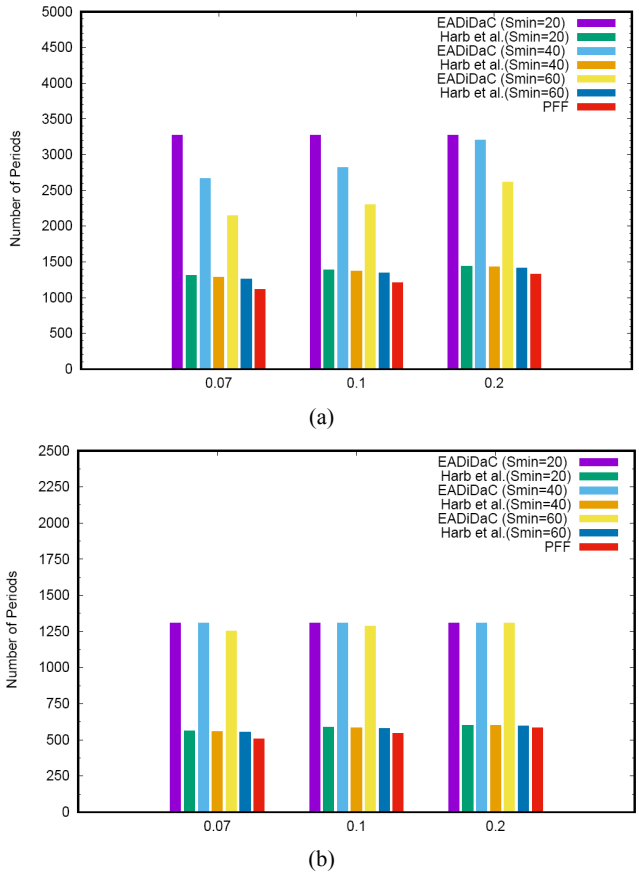
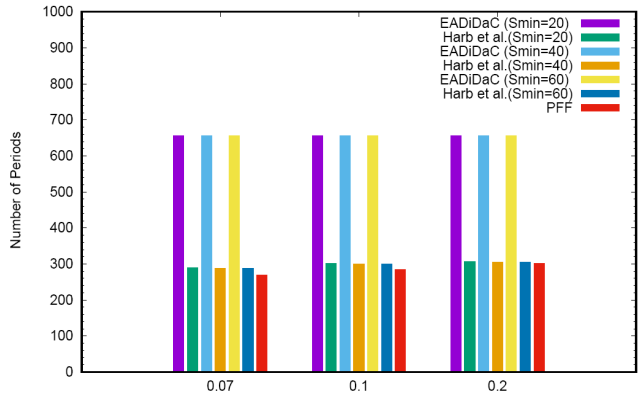
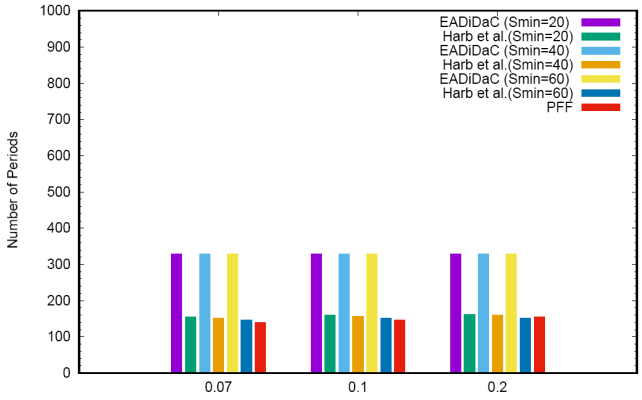


Figure 14 Lifetime of a sensor node, (a) $\rho = 20$ (b) $\rho = 50$ (c) $\rho = 100$ (d) $\rho = 200$ (continued) (see online version for colours)



(c)



(d)

4.3.5 Algorithmic complexity and t-test

As an analytical study, every sensor node n_i constructs sequence of sensed data S_i of ρ temperature readings. The time complexity of the Algorithm 1 is $O(\rho \times \log_2(\rho))$. The time complexity of the Algorithm 2 is O(m|SG|), where m is the number of segments and |SG| is the length of the segment. Algorithm 3 has O(m) as a computation complexity. The time complexity of the Algorithm 4 is O(|Q| × |T|), where |Q| and |T| are the number of segments for data series Q and T respectively. Therefore, the time complexity of our proposed EADiDaC method in the worst case is O(|Q| × |T|) and it will save at most $(2 \times \rho)$ measures at the memory of the sensor node in each period. Therefore, the storage (space) complexity of EADiDaC method is $O(\rho)$. The time complexity of Harb et al. algorithm takes $O(\rho^2)$. Finally, the time complexity of PFF is $O(\rho \times \log_2(\rho))$. In addition, the complexity of the message in EADiDaC method depends mainly on the number of collected data (ρ) in the period, which is fixed by the application. If it is required a large value for ρ, several solutions can be used such as data packet division. The space

complexity depends on the sensor node memory size as well as ρ, which can be handled in a similar way to the complexity of the message.

In addition, we use the statistical analysis such as t-test to show that our results are significant. Therefore, the t-test is applied on the comparison result of the energy consumption between our proposed EADiDaC method and the two existing methods (Harb et al. and PFF). The t-test (with p-value) between EADiDaC and Harb et al. is equal to 1.18254E-10, whilst the t-test (with p-value) between EADiDaC and PFF is equal to 9.89085E20. Hence, the t-test (with p-value < 0.05) shows that our result is significant and the energy consumption is significantly reduced.

5 Conclusions and future works

This paper presents a method, called distributed adaptive data collection method (EADiDaC), which collects periodically sensor readings and improves the PSN lifetime. EADiDaC method works into cycles and consists of four phases. First, collecting the data readings. Second, the sensor converts the collected temperature readings into APCA representation in order to reduce its dimensionality. Third, the redundant collected readings are reduced using SAX approach. Fourth, sampling resolution to adapt the rate of sampling at the sensor node in accordance with the dynamic changing of observed environment. EADiDaC method considers the risk level of an application by fixing the minimum sampling rate that permits to sensor node to collect readings at a minimum rate while maintaining a good quality of the collected readings. To assess the effectiveness of EADiDaC method, we compared it with two other methods using several performance metrics like a number of collected and sent readings, energy consumption, and PSN lifetime. Simulation results show the efficiency of EADiDaC method to conserve the energy at the sensor nodes thus prolong the PSN lifetime.

In future, we plan to improve our work to consider the sensing overlap among sensor nodes at the aggregator level to optimise both the aggregated readings and lifetime while maintaining a good accuracy.

References

Abdelaal, M., Theel, O., Kuka, C., Zhang, P., Gao, Y., Bashlovkina, V., Nicklas, D. and Fränzle, M. (2016) 'Improving energy efficiency in QoS-constrained wireless sensor networks', *International Journal of Distributed Sensor Networks*, Vol. 12, No. 5, p.1576038.

Akyildiz, I.F. and Vuran, M.C. (2010) *Wireless Sensor Networks*, Vol. 4, John Wiley & Sons, UK.

Alippi, C., Anastasi, G., Di Francesco, M. and Roveri, M. (2010) 'An adaptive sampling algorithm for effective energy management in wireless sensor networks with energy-hungry sensors', *IEEE Transactions on Instrumentation and Measurement*, Vol. 59, No. 2, pp.335–344.

Bahi, J.M., Makhoul, A. and Medlej, M. (2014) 'A two tiers data aggregation scheme for periodic sensor networks', *Adhoc & Sensor Wireless Networks*, Vol. 21, No. 1, pp.77–100.

Campobello, G., Segreto, A. and Serrano, S. (2016) 'Data gathering techniques for wireless sensor networks: a comparison', *International Journal of Distributed Sensor Networks*, Vol. 12, No. 3, p.31.

Cassisi, C., Montalto, P., Aliotta, M., Cannata, A. and Pulvirenti, A. (2012) 'Similarity measures and dimensionality reduction techniques for time series data mining', *Advances in Data Mining Knowledge Discovery and Applications*, InTech, Rijeka.

Chakrabarti, K., Keogh, E., Mehrotra, S. and Pazzani, M. (2002) 'Locally adaptive dimensionality reduction for indexing large time series databases', *ACM Transactions on Database Systems (TODS)*, Vol. 27, No. 2, pp.188–228.

Chatterjea, S. and Havinga, P. (2008) 'An adaptive and autonomous sensor sampling frequency control scheme for energy-efficient data acquisition in wireless sensor networks', *International Conference on Distributed Computing in Sensor Systems*, Springer, Berlin Heidelberg, pp.60–78.

de Graaf, M. (2013) 'Energy efficient networking via dynamic relay node selection in wireless networks', *Computer Communications*, Vol. 11, No. 3, pp.1193–1201.

Ganz, F., Barnaghi, P. and Carrez, F. (2014) 'Multiresolution data communication in wireless sensor networks', *IEEE World Forum on Internet of Things (WF-IoT)*, IEEE, pp.571–574.

Gedik, B., Liu, L. and Philip, S.Y. (2007) 'ASAP: an adaptive sampling approach to data collection in sensor networks', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 12, pp.1766–1783.

Gupta, S.K. (2010) *Routing and Data Dissemination in Wireless Sensor Networks*, Doctoral dissertation, National Institute of Technology, Kurukshetra-136119, India.

Harb, H., Makhoul, A., Jaber, A., Tawil, R. and Bazzi, O. (2016) 'Adaptive data collection approach based on sets similarity function for saving energy in periodic sensor networks', *International Journal of Information Technology and Management*, Vol. 15, No. 4, pp.346–363.

Huang, J., Qian, F., Guo, Y., Zhou, Y., Xu, Q., Mao, Z.M., Sen, S. and Spatscheck, O. (2013) 'An in-depth study of LTE: effect of network protocol and application behavior on performance', *ACM SIGCOMM Computer Communication Review*, Vol. 43, No. 4, pp.363–374.

Idrees, A.K., Deschinkel, K., Salomon, M. and Couturier, R. (2014) 'Coverage and lifetime optimization in heterogeneous energy wireless sensor networks', *ICN 2014*, p.60.

Idrees, A.K., Deschinkel, K., Salomon, M. and Couturier, R. (2015) 'Distributed lifetime coverage optimization protocol in wireless sensor networks', *The Journal of Supercomputing*, Vol. 71, No. 12, pp.4578–4593.

Idrees, A.K., Deschinkel, K., Salomon, M. and Couturier, R. (2016) 'Perimeter-based coverage optimization to improve lifetime in wireless sensor networks', *Engineering Optimization*, Vol. 48, No. 11, pp.1951–1972.

Jain, A. and Chang, E.Y. (2004) 'Adaptive sampling for sensor networks', *Proceedings of the 1st International Workshop on Data Management for Sensor Networks: In Conjunction with VLDB 2004*, ACM, pp.10–16.

Jain, A., Chang, E.Y. and Wang, Y.F. (2004) 'Adaptive stream resource management using kalman filters', *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, ACM, pp.11–22.

Jon, Y. (2016) *Adaptive Sampling in Wireless Sensor Networks for Air Monitoring System*, Dissertation [online] http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-295995.

Jun, Z., Simplot-Ryl, D., Bisdikian, C. and Mouftah, H.T. (2011) 'The internet of things', *IEEE Commun. Mag*, Vol. 49, No. 11, pp.30–31.

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

Law, Y.W., Chatterjea, S., Jin, J., Hanselmann, T. and Palaniswami, M. (2009) 'Energy-efficient data acquisition by adaptive sampling for wireless sensor networks', *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, ACM, pp.1146–1151.

Layuan, L., Chunlin, L. and Peiyan, Y. (2007) 'Performance evaluation and simulations of routing protocols in ad hoc networks', *Computer Communications*, Vol. 30, No. 8, pp.1890–1898.

Lazaridis, I. and Mehrotra, S. (2003) 'Capturing sensorgenerated time series with quality guarantees', *19th International Conference on In Data Engineering, Proceedings*, IEEE, pp.429–440.

Le Borgne, Y.A., Santini, S. and Bontempi, G. (2007) 'Adaptive model selection for time series prediction in wireless sensor networks', *Signal Processing*, Vol. 87, No. 12, pp.3010–3020.

Li, S., Da Xu, L. and Wang, X. (2013) 'Compressed sensing signal and data acquisition in wireless sensor networks and internet of things', *IEEE Transactions on Industrial Informatics*, Vol. 9, No. 4, pp.2177–2186.

Li, Y., Ye, W., Heidemann, J. and Kulkarni, R. (2008) 'Design and evaluation of network recon\uration protocols for mostly-off sensor networks', *Ad Hoc Networks*, Vol. 6, No. 8, pp.1301–1315.

Lin, J. and Li, Y. (2009) 'Finding structural similarity in time series data using bag-of-patterns representation', *International Conference on Scientific and Statistical Database Management*, pp.461–477, Springer, Berlin Heidelberg.

Liu, C., Wu, K. and Pei, J. (2007) 'An energy efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 7, pp.1010–1023.

Liu, C., Wu, K. and Tsao, M. (2005) 'Energy efficient information collection with the ARIMA model in wireless sensor networks', *IEEE Global Telecommunications Conference on GLOBECOM'05*, IEEE, Vol. 5, 5pp.

Madden, S. (2004) *Intel Berkeley Research Lab Data* [online] http://berkeley.intel-research.net/labdata (accessed 25 August 2016).

Makhoul, A., Harb, H. and Laiymani, D. (2015) 'Residual energy-based adaptive data collection approach for periodic sensor networks', *Ad Hoc Networks*, Vol. 35, No. C, pp.149–160.

Malinowski, S., Guyet, T., Quiniou, R. and Tavenard, R. (2013) '1d-sax: a novel symbolic representation for time series', *International Symposium on Intelligent Data Analysis*, Springer, Berlin-Heidelberg, pp.273–284.

Masoum, A., Meratnia, N. and Havinga, P.J. (2012) 'A decentralized quality aware adaptive sampling strategy in wireless sensor networks', *9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC)*, IEEE, pp.298–305.

Masoum, A., Meratnia, N. and Havinga, P.J. (2013) 'An energy-efficient adaptive sampling scheme for wireless sensor networks', *IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, IEEE, pp.231–236.

Mehrnoush, M., Fathi, R. and Vakili, Vahid, T. (2015) 'Proactive spectrum handoff protocol for cognitive radio ad hoc network and analytical evaluation', *IET Communications*, Vol. 9, No. 15, pp.1877–1884.

Mohsenifard, E. and Ghaffari, A. (2016) 'Data aggregation tree structure in wireless sensor networks using cuckoo optimization algorithm', *Information Systems & Telecommunication*, Vol. 4, No. 3, pp.182–190.

Nam, Y., Lee, H., Jung, H., Kwon, T. and Choi, Y. (2006) 'An adaptive MAC (A-MAC) protocol guaranteeing network lifetime for wireless sensor networks', *12th European in Wireless Conference 2006 – Enabling Technologies for Wireless Multimedia Communications (European Wireless)*, VDE, pp.1–7.

Padhy, P., Dash, R.K., Martinez, K. and Jennings, N.R. (2010) 'A utility-based adaptive sensing and multihop communication protocol for wireless sensor networks', *ACM Transactions on Sensor Networks (TOSN)*, Vol. 6, No. 3, p.27.

Srbinovski, B., Magno, M., O'Flynn, B., Pakrashi, V. and Popovici, E. (2015) 'Energy aware adaptive sampling algorithm for energy harvesting wireless sensor networks', *Sensors Applications Symposium (SAS), IEEE*, pp.1–6.

Tang, X. and Xu, J. (2008) 'Adaptive data collection strategies for lifetime-constrained wireless sensor networks', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 19, No. 6, pp.721–734.

Tulone, D. and Madden, S. (2006) 'PAQ: time series forecasting for approximate query answering in sensor networks', *European Workshop on Wireless Sensor Networks*, Springer, Berlin Heidelberg, pp.21–37.

Ulusoy, A., Gurbuz, O. and Onat, A. (2011) 'Wireless model-based predictive networked control system over cooperative wireless network', *IEEE Transactions on Industrial Informatics*, Vol. 7, No. 1, pp.41–51.

Van Dam, T. and Langendoen, K. (2003) 'An adaptive energy-efficient MAC protocol for wireless sensor networks', *Proceedings of the 1st international Conference on Embedded Networked Sensor Systems*, ACM, pp.171–180.

Varga, A. (2003) *OMNeT++ Discrete Event Simulator* [online] https://omnetpp.org/ (accessed 20 August 2016).

Wang, C., Ma, H., He, Y. and Xiong, S. (2012a) 'Adaptive approximate data collection for wireless sensor networks', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 23, No. 6, pp.1004–1016.

Wang, J., Tang, S., Yin, B. and Li, X.Y. (2012b) 'Data gathering in wireless sensor networks through intelligent compressive sensing', *Proceedings IEEE in INFOCOM*, IEEE, pp.603–611.

Wang, Y., Wang, P., Pei, J., Wang, W. and Huang, S. (2013) 'A data-adaptive and dynamic segmentation index for whole matching on time series', *Proceedings of the VLDB Endowment*, Vol. 6, No. 10, pp.793–804.

Willett, R., Martin, A. and Nowak, R. (2004) 'Backcasting: adaptive sampling for sensor networks', *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, ACM, pp.124–133.

Yahmed, Y.B., Bakar, A.A., Hamdan, A.R., Ahmed, A. and Abdullah, S.M.S. (2015) 'Adaptive sliding window algorithm for weather data segmentation', *Journal of Theoretical and Applied Information Technology*, Vol. 80, No. 2, p.322.

Ye, W., Heidemann, J. and Estrin, D. (2002) 'An energy-efficient MAC protocol for wireless sensor networks', *Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE in INFOCOM 2002*, IEEE, Vol. 3, pp.1567–1576.

Zhang, J., Ren, L., Ding, Y. and Hao, K. (2015) 'Adaptive sampling algorithm with endocrine regulation mechanism for wireless sensor network', *10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, IEEE, pp.502–507.

Zheng, T., Radhakrishnan, S. and Sarangan, V. (2005) 'PMAC: an adaptive energy-efficient MAC protocol for wireless sensor networks', *19th IEEE International Parallel and Distributed Processing Symposium*, IEEE, 8pp.

Zifan, A., Moradi, M.H., Saberi, S. and Towhidkhah, F. (2007) 'Automated segmentation of ECG signals using piecewise derivative dynamic time warping', *International Journal of Biological and Life Sciences*, Vol. 2, No. 8, pp.181–185.

Notes

1 The other is done by the same manner.